Grenze International Journal of Engineering and Technology, June Issue

Breast Cancer Survival Prediction using Gene Expression Data

Arun Sebastian[1] and Dr. Lija Jacob[2]
[1-2]Christ University, Pune Lavasa Campus, Maharashtra, India
Email: arun.sebastian@science.christuniversity.in, lija.jacob@christuniversity.in

***Abstract*—Breast cancer is one of the most common forms of cancer in the world.[1]. Breast, skin, colon, pancreatic, and other 100 types of cancer have founded globally. An accurate breast cancer prognosis can save many patients from having unnecessary treatment and the huge medical costs that come with it. Multiple gene mutations can possibly transform a normal cell into a cancerous one. Genomic variations and traits have a significant effect on cancer. Genetic abnormalities caused by various circumstances drive numerous efforts to find biomarkers of breast cancer advancement. Early Detection of Cancer types is the only way to recover the patients from this acute disease. In this paper, a proposed Deep learning algorithm and Machine learning algorithms are used to predict the survival of cancer patients using clinical data and gene expression data. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset is split into clinical and gene data for detailed pre-processing. This proposed method gives a better understanding of the condition and assesses how effective treatment methods are by using Deep Learning and Machine Learning models on gene data. Logistic Regression is the most accurate method identified.**

***Index Terms*— Breast cancer, Deep Learning, Machine Learning, Dimensionality Reduction, Gene Expression.**

I. INTRODUCTION

Breast Cancers are increasing day by day due to mutations in the genes and lifestyle of people. It's essential to increase the survival rate by efficient treatment methods. Mammographic screening has resulted in the early discovery of tumors, even small non-invasive tumors, and breakthroughs in therapy, lowering cancer mortality rates. On the other hand, technological advancements and the advent of precision medicine lead to a shift away from "one size fits all" breast cancer screening. Early cancer genomes detection can help choose a therapy strategy. Treatment-resistant subclones are more likely to be found in more diverse tumors. Resistant subclones survive and multiply, forming a mixed tumor. According to the World Health Organization, cancer has surpassed heart disease as the second biggest cause of mortality worldwide, accounting for nearly 1 in every six deaths.
Furthermore, malignancies are highly diverse, with people with comparable diagnoses receiving the same treatment regimen experiencing vastly different results. This has prompted academics to look for more indicators to aid in predicting individual outcomes. Many of these studies rely solely on clinical data. Unfortunately, whereas lymph node status and histological grade are predictive of metastases, they do not appear to be enough to classify clinical outcomes correctly. The correct prognosis of a disease outcome, on the other hand, is one of the most exciting and complex challenges for clinicians [2]. ML approaches have become a prominent tool for

Grenze ID: 01.GIJET.8.2.13

medical researchers due to this. These methods can extract patterns and relationships from large datasets and accurately forecast future cancer outcomes. Mutations cause hereditary breast cancer in the BRCA1 or BRCA2 genes in many cases. These malignancies often have histological abnormalities that are indicative of a mutant gene. They predicted that the genes expressed by these two types of tumours would be unique, allowing gene-expression profiles to be used to identify cases of hereditary breast cancer [3] [4].
Patients with breast cancer at the same stage can have profoundly varying treatment responses and outcomes. The most effective biomarkers of metastasis (such as lymph node status and histological grade) fail to distinguish breast tumors accurately based on their clinical behaviour [5]. Chemotherapy or hormone therapy reduces the risk of distant metastases by around a third, 70–80% of patients who receive it would have survived without it [6].
The overall sections of paper are organized as follows. The second section discuss the related works. The third section will discuss data acquisition and pre-processing. The fourth section analyse the treatment types and survivals. In the fifth section demonstrate Dimensionality reduction method. In the sixth section explains the algorithms used in the paper. In the seventh section discuss the results obtained from the experiments done in the paper and its discussion. In the final section, summarize the accomplished and future work.

II. Related Works

Many different machine learning and deep learning methods are covered in various research papers. Artificial neural networks have effectively applied in various clinical settings for pattern recognition and survival prediction. [7].
In [8], Yingshuai Sun1 et al. proposed a new model GDL (genome Deep Learning), which includes a DNN (Deep Neural Network) model for cancer identification based on genome variation. The authors have identified 12 Cancer types using DNN within a transform framework. With this advanced implementation of a sizeable genomic sequence model, the cancers can be located even in the earliest stage. As the system uses DNN models with several computation layers to identify the cancer types which a multiclass classification arrangement, the accuracy, sensitivity, and specificity are used to evaluate the performance of the classifiers. All individual models had an accuracy rate ranging from 97.47%(PRAD) to 100% (KIRC, LUSC, OV). The entire specific model's accuracy, sensitivity, and specificity were 97.70%, 94.30%, and 85.54%, respectively. To confirm the relationship between the number of standard dimensions and cancer kinds, a statistical study was performed.
In [9], Padideh danaee et al. Proposed a deep learning approach to cancer detection and the identification of genes critical for diagnosing breast cancer. They propose a deep learning technique to cancer detection and the identification of breast cancer-related genes. SDAE (Stacked Denoising Autoencoder) identifies functional information from high-dimensional gene expression profiles and converts them to a more comprehensible, lower-dimensional representation. SDAEs are used to improve cancer detection by extracting both linear and nonlinear correlations in the data and, perhaps more importantly, extracting a subset of relevant genes from deep models as a set of prospective cancer biomarkers. Principal Component Analysis was performed to cluster the important genes in the context of expression profiles due to the large number of genes and the complexity of the data. Out of the five measures we looked at, SDAE features performed the best. The ML model with the SDAE feature has the highest accuracy, 98.73 percent [10].
In [11], Aaron Mckenna et al. implement a Genome Analysis Toolkit (GATK) is a structural programming framework that uses the MapReduce functional programming philosophy to make it easier to create efficient and reliable analysis tools for next-generation DNA sequencers. GATK can also handle advanced features like distributed and automatic shared-memory parallelization thanks to its software core. The Nave genotyper does a good job of detecting variations — on chromosome 1, 315,202 variants were called, with 81.70 percent in dbSNP and a concordance accuracy of 99.76 percent.
There is an increase in the articles on the topic applicability of Machine Learning (ML) techniques and the Polygenic Risk Score (PRS) for predicting genetic diseases. PRS is a traditional approach by using a validation sample. The PRS can be calculated as a sum of the alleles associated with the phenotype (often weighted by the specific SNP (Single Nucleotide polymorphisms) coefficients of the GWAS (Genome-Wide Association Studies.)). Using this score, multiple SNPs with the given trait can be evaluated. ML approaches disease prediction, disease diagnosis, drug development, detect epitasis within the human genome [12]. The Pan-Cancer Atlas' cancer classification uses ribonucleic acid sequencing (RNA-seq) to convert 1-dimensional (1D) gene expression levels into 2-dimensional (2D) images. The input picture (logit of the input images) provides a map of the same image with the relevant salient area (or chunk of the image) for predicting the gene in the CNN at

the conclusion of the convolutional layer. The heatmap aids visualization, and the intensity created represents each gene's score in the classification task.[13].
A model is classified based on the class membership of its nearest neighbors in the gene space. It combines a genetic algorithm (GA) and the k-nearest neighbor (KNN) method to identify genes that jointly can discriminate between two types of samples (e.g., standard vs. tumor). The dimensionality of the gene subspace is arbitrarily set to 50. Intuitively, samples (objects) become more distinct (dissimilar) when more genes are compared. Once such a set of relevant genes has been identified, an unknown sample can be classified by comparing its expression profile with the known samples. GA is applied in combination with the KNN method to determine the many subsets of genes that can discriminate tumors from standard tissue samples. [14].

III. Dataset and Preprocessing

A. Dataset

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset is used in this paper. The dataset contains Clinical features, m-RNA levels z-score, and genes mutations for 1904 patients. The genetics section of the collection includes 331 genes with z-scores for m-RNA levels and 175 genes with mutations. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database is a joint project between Canada and the United Kingdom that comprises targeted sequencing data from 1,980 primary breast cancer samples [15].

B. Preprocessing

The dataset contains both clinical features as well as gene mutations. It is difficult to process with the whole dataset so the first step is to split the dataset as Clinical and Gene mutations separately. Now the preprocessing steps can be done easily with respective datasets.
Null Values: Missing data means the values that aren't recorded in a dataset. They can range from a single value missing in a single cell to the complete loss of an observation (row). Missing datacan arise in both continuous and categorical variables. Clinical dataset occurs missing values which fill with mode for categorical variables. Also, deals the nominal and ordinal data with one-hot encoding and label encoding. Finally join those encoded dataset into a new one.
Data Cleaning: Data cleaning includes removing unwanted data. Dropping some features and arrange the dataset in a way to build the model appropriately. Rounding of the age of patients will be better to know which age category has more number of cancer patients. Drop some irrelevant columns such as: patient_id and cancer_type.
Correlation: In order to make better understanding and make some inferences out of the data the correlation helps a lot. It is a data visualization technique which enables to bring out more conclusions. Correlation plot has used to find out which all features are varying with target column in other ways we are looking into the the correlation between the variables. According to the high positive correlation and negative correlation the features are been chosed. Some of the variables has high correlation while some of them has less.
Data redundancy: Important reason to search for data redundancy is that having the same data in multiple places might confuse users and make it difficult for them to figure out which datathey should access or edit. In the case of Breast cancer dataset if we have the duplicate data of some patients, it is meaningless to find the survival detection or cancer detection from those dataset.
Feature Scaling: Feature scaling is a significant stage in the pre-processing of data before building a machine learning model. The difference between a bad and a good machine learning model can be determined by scaling. Normalization and Standardization are the most frequent feature scaling approaches. Genomic data is ranging from very low to high. To scale those features into a particular range MinMax scaler has used.

IV. Treatment Types and Survivals

There are many treatment methods are innovating day by day for breast cancer but the highest recommended methods are Chemotherapy, Hormone Therapy, Radio Therapy, Surgical removal, Medication and so on. Some patients require the combination of these treatment methods according to the severity of the patients' stage. In this paper three types of treatment methods are mainly discussing Chemotherapy, Hormone Therapy and Radio Therapy. The patients who undergo the treatment and out of them how many are able to survived and died even after taking this treatment. In the below (Figure 1.) it is clearly visualized that the patients who survived and died after taking these treatments. Radio therapy has high number of survived patients 520 while Chemotherapy shows the smaller number of survived patients. Chemotherapy patients have high number of death rate compared

to other treatment methods. Hormone therapy results the second highest number of survived patients. From this it gives a conclusion that separately radio therapy is the best treatment method among this three while chemotherapy performs less.

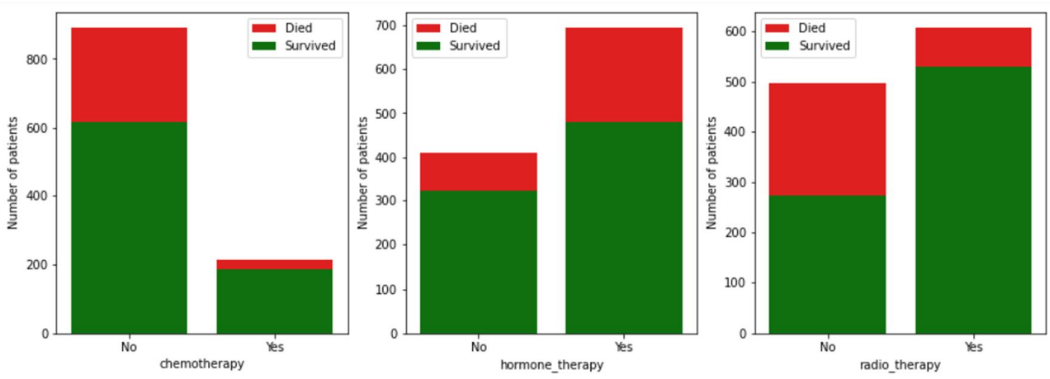

Figure 1 The distribution of treatment and survival

The correlation among these three treatment types chemotherapy and hormonal depicts the highest negative correlation where radio and hormone have less correlation. The below (Figure 2) explains how the combination of treatment methods effectively results in patients' survival rate.

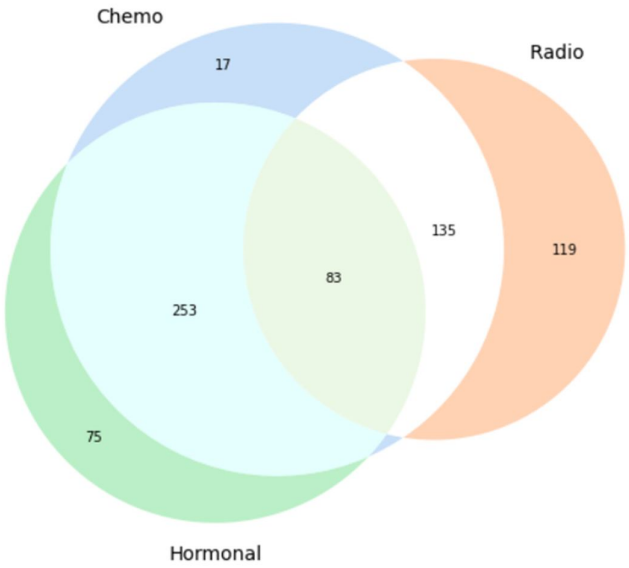

Figure. 2. Survived patients by treatment group

It depicts the same results which get in the correlation as the combination of chemotherapy and hormonal therapic patients survived most with 253 while those who taken only chemotherapy has a smaller number of survived patients 17. The combination of radio therapy and hormone therapic patients has the second highest survived patients 135.So according to the patient's cancer stage this combined treatment effectively impact on its cure.

V. Dimensionality Reduction with Principal Component Analysis

Principal Component Analysis is a reduction technique we utilized (PCA). This method reduces a collection of strongly correlated variables to a set of principal components, which can subsequently be used to make a decision. High dimensionality limits models from producing effective results since it is difficult for the model to

extract target information from the data. In order for models to work, their dimensionality must be minimized. The variance in the data is represented by these primary components. Each succeeding primary component represents less variance in the data [8]. In this paper as there are a greater number of gene combinations it is difficult to take all 695 genes in model implementation. Most of them are not important genes which is less common, and most of them do not increase the risk of breast cancer. So, the dimensionality reduction technique is to reduce the combination of genes. Principal Component Analysis method is used to reduce the dimensions.

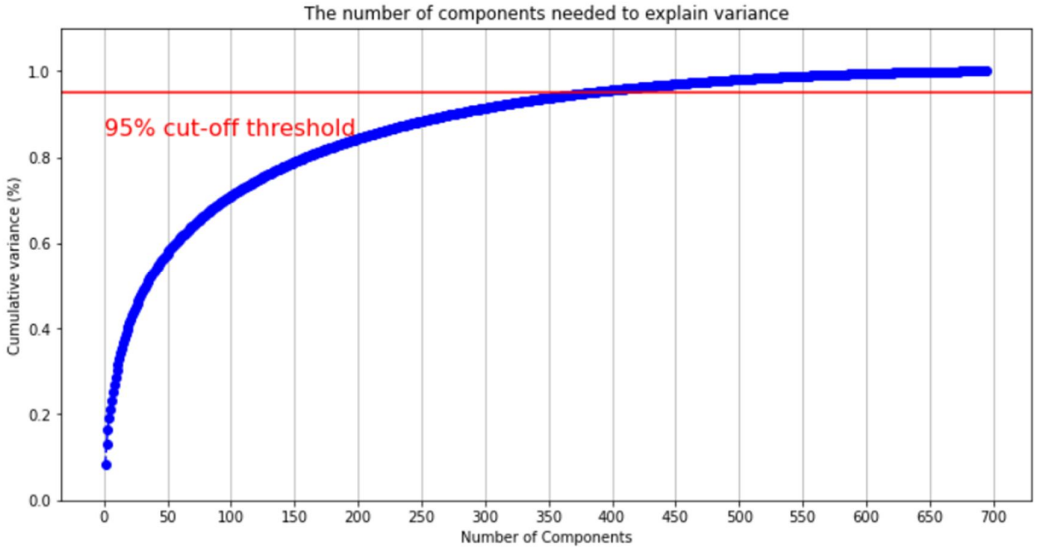

Figure.3. PCA Cumulative explained variance

In order to pick how many components to shrink the data cumulative explained variance in (Figure.3) shows the actual number of components need to be taken. From this genomic dataset exactly 375 components only needed for further model building. Taking 695 components will results in overfitting also the trained model achieves nearly 100% accuracies. To avoid this problem PCA is the apt dimensionality reduction method to select the best features among the genomic dataset.

VI. ALGORITHMS USED

Identified the optimum model with accuracy by combining several machine learning and deep learning models [17]. In Deep Learning Artificial Neural Network and In Machine Learning Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Extra Trees, Adaboost and RandomForestRegressor are some of the models that is used in this paper. KNN and Logistic Regression are the best among these proposed models.

A. Artificial Neural Network

Artificial Neural Networks (ANNs) are a type of computational learning system that employs a network of functions to comprehend and translate a data input in one form into a desired output in another. Neurons in neural networks are linked to one another in various layers of the network.

$$\sum wixi + bias = w1x1 + w2x2 + w3x3 + bias$$

Once an input layer is determined, weights are assigned. After that, all of the inputs are multiplied by their respective weights and then combined together. The output is then passed through an activation function to determine the output.

B. K – Nearest Neighbor Classifier

The method saves all available cases (test data) and categorizes new cases based on the majority votes of its K neighbors. The initial step in developing KNN is to convert data points into mathematical values (vectors). The technique works by calculating the distance between these points' mathematical values. It finds the chance of the points being similar to the test data by computing the distance between each data point and the test data. The

probabilities of which points share the highest probabilities are used to classify them. The procedure will determine the k-nearest neighbors of the data point for a given value of K, and then assign the class to the data point based on which class has the h largest number of data points among all classes of the K neighbors.

C. Logistic Regression

Logistic regression is a "supervised machine learning" approach that can be used to model the likelihood of a specific class or occurrence. When the data is linearly separable and the result is binary or dichotomous, this method is applied. As a result, for Binary classification tasks, Logistic regression is commonly utilized. Predicting an output variable that is discrete in two classes is referred to as binary classification. When the 'multi class' option is set to'ovr,' the training algorithm utilises the one-vs-rest (OvR) scheme, and when the 'multi class' option is set to 'multinomial,' the training algorithm uses the cross-entropy loss.

$$\text{Ln}\left(\frac{P}{1-P}\right)=\beta_O+\beta_1X_1+\beta_2X_2+\ldots\beta_kX_k$$

Thus, the value lies between -1 and +1. If the value of the logit function lies below 0, then it is of class 0 and if the value of the logit function lies above 0, then it is of class 1.

VII. Results and Discussion

A. Deep Learning and Machine Learning

In this paper the different models experiment on the combined clinical data and gene expression data. By doing PCA the selected genes and clinical features are finally taking after various levels of preprocessing steps. Artificial Neural Network (ANN) is used as a Deep Learning approach which is fully connected layers such as input layer and output layer with two hidden layers. Input layer and hidden layers are having Relu activation function and SoftMax function as the activation function as the output layer. Thus, accuracy obtained 57.50.

In this project the model train with 80% of the total data space andused 20% for model evaluation. The models are trained and tested using the new dataset that has made after extracting unwanted features from both gene data and clinical data. The models used in Machine Learning are Logistic Regression, KNN, Extra tree, Random Forest Regressor, Random Forest, Adaboost and Decision Tree. Minimum Covariance Determinant the rationale for choosing this model is that the problem statement is a regression or a survival detection problem and among the set of models for this task the mentioned model performed very much well for the given task of classification or survival detection of Breast Cancer. For Machine Learning models import the library grid search CV and validation parameters. Then define two functions such as hyper tuning and model metrics. In Hyperparameter tuning specifies a new model and include several parameters into it. Then while building the model in that two functions are passed to get better accuracy for each machine learning models. After defining the class, every machine learning models implemented in the single cell with learning rate parameters.

Confusion matrix, F1-Score and Precision recall are performance metrics for machine learning models. The F1-Score is theharmonic mean of Recall and Precision, and it provides a more accurate picture of cases that were mistakenly classified than the Accuracy Metric. Even though most models do not give an ideal F1-Score, the system performs well for the task at hand. The final system prediction is dependent on predictions made by models of both supervised and unsupervised learning. In the below (Table 1.) it depicts that among these machine learning models Logistic Regression is obtained the highest accuracy 0.8320 followed by K Nearest Neighbors 0.7191, Decision Tree 0.6614, Adaboost 0.6351, Random Forest 0.5958, ANN 0.5750, Random Forest Regressor 0.5727, Extra Trees 0.5590.

TABLE I. MACHINE LEARNING AND DEEP LEARNING ALGORITHM WITH ACCURACY

Algorithm	**Accuracy**
Logistic Regression	0.8320
K Nearest Neighbors	0.7191
Decision Tree	0.6614
Adaboost	0.6351
Random Forest	0.5958
ANN	0.5750
Random Forest Regressor	0.5727
Extra Trees	0.5590

B. Discussion

In this paper try to identify the survival rate of breast cancer patients by using gene expression data with clinical data features. Breast cancer is a form of cancer that develops when certain cells in the breast become abnormally large. This form of cancer is responsible for the majority of cancer-related deaths in women. Experiments done on Machine Learning and Deep Learning algorithms. Among Machine Learning and Deep Learning model Machine Learning model shows the best accuracy. Deep into it in Machine Learning model Logistic Regression shows the best accuracy that is 83.20%. while Extra Trees has the less accuracy with 55.90%. The accuracy is less even for the best model is that the gene types are more and it's difficult to find the relation between the gene and clinical data and to predict the survival rate. In this project the model train with 80% of the total data space andused 20% for model evaluation. The models are trained and tested using the new dataset that has made after extracting unwanted features from both gene data and clinical data.

Combination of gene expression data and clinical data has given a good mileage for prediction of cancer patients survival. Also, this project put light through the treatment methods. By analyzing it have been found that combination of chemotherapy and hormonal treatment patients are survived more. However, the patients did only chemotherapy has the less survival number of patients.

VIII. CONCLUSION

As the number of Breast Cancers are increasing day by day the challenges for medical practitioners and medical scientists are also increasing proportionally. Gene Expression has a great impact on developing a breast cancer disease in humans. Gene data add more accurate prediction because the breast cancers are mostly transfer due to this cancerous gene mainlyBRCA1 and BRCA2.In this paper METABRIC dataset has used for prediction. While analyzing the treatment methods combination of Chemotherapy and Hormonal Therapy results highest number of survival rates. It is clear that combination of best treatment methods will cure the patients according to their stage of cancer. Deep Learning and Machine Learning algorithms are used to predict the survival rate of breast cancer patients. Among these Logistic Regression shows the better accuracy of 83%. In future we can relate the gene, by which cancer type is associated with it. So, we can identify that particular gene. If we get to know that particular gene by multiple cell divisions or any other advanced medical technology, we can suppress that gene to prevent this diseasefrom next generation.

REFERENCES

[1] Lin X, Liu C, Sheng Z, Gong X, Song L, Zhang R, Zheng H, Sun M. "Highly Sensitive Fluorescence and Photoacoustic Detection of Metastatic Breast Cancer in Mice Using Dual-Modal Nanoprobes". ACS Appl Mater Interfaces. 2018 Aug 8;10(31): 26064-26074.doi: 10.1021/acsami.8b09142. Epub 2018 Jul 25. PMID: 30044603.

[2] Pramanik, Sabyasachi & Sagayam, Martin & Jena, Om. "Machine Learning Frameworks in Cancer Detection." E3S Web of Conferences. 297. 01073. 10.1051/e3sconf/202129701073, 2021

[3] Hedenfalk, Ingrid & Duggan, David & Chen, Yidong & Radmacher, Michael & Bittner, Michael & Simon, Richard & Meltzer, Paul & Gusterson, Barry & Esteller, Manel & Kallioniemi, Olli & Wilfond, Benjamin & Borg, Ake & Trent, Jeffrey & Raffeld, Mark & Yakhini, Zohar & Ben-Dor, Amir & Dougherty, Edward & Kononen, Juha & Sauter, Guido. "Gene-Expression Profiles in Hereditary Breast Cancer." The New England journal of medicine. 344. 539-48. 10.1056/NEJM200102223440801,2001.

[4] Yang, Haiyan & Yu, Li-Rong & Yi, Ming & Lucas, David & Lukes, Luanne & Lancaster, Mindy & Chan, King & Issaq, Haleem & Stephens, Robert & Conrads, Thomas & Veenstra, Timothy & Hunter, Kent. (2006). "Parallel Analysis of Transcript and Translation Profiles: Identification of Metastasis-Related Signal Pathways Differentially Regulated by Drug and Genetic Modifications". Journal of proteome research. 5. 1555-67. doi:10.1021/pr0504283.

[5] Veer, Laura J.; Dai, Hongyue; van de Vijver, Marc J.; He, Yudong D.; Hart, Augustinus A. M.; Mao, Mao; Peterse, Hans L.; van der Kooy, Karin; Marton, Matthew J.; Witteveen, Anke T. "Gene expression profiling predicts clinical outcome of breast cancer", Nature 415(6871):530-6, DOI:10.1038/415530a, February 2002

[6] Suzanne A Eccles; Danny R Welch. "Metastasis: recent discoveries and novel treatment strategies", doi: 10.1016/S0140-6736(07)60781-8, PMC 2008 May 19

[7] W.G Baxt (1995). "Application of artificial neural networks to clinical medicine", 346(8983), 1135–1138. doi:10.1016/s0140-6736(95)91804-3, 28 October 1995

[8] Sun, Yingshuai, Sitao Zhu, Kailong Ma, Weiqing Liu, Yaojing Yue, Gang Hu, Huifang Lu and Wenbin Chen. "Identification of 12 cancer types through genome deep learning." Sci Rep 9, 17256 (2019). https://doi.org/10.1038/s41598-019-539893

[9] Danaee, Padideh, Reza Ghaeini and D. Hendrix. "A Deep Learning Approach for Cancer Detection and Relevant Gene Identification." Pacific Symposium on Biocomputing. 2017, DOI: 10.1142/9789813207813_0022

[10] M. K. Elbashir, M. Ezz, M. Mohammed and S. S. Saloum, "Lightweight Convolutional Neural Network for Breast

Cancer Classification Using RNA-Seq Gene Expression Data," in IEEE Access, vol. 7, pp. 185338-185348, 2019, doi: 10.1109/ACCESS.2019.2960722.

[11] McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. DePristo. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research 20 9 (2010): 1297-303.

[12] N. Mena Mamani, "Machine Learning techniques and Polygenic Risk Score application to prediction genetic diseases", ADCAIJ, vol. 9, no. 1, pp. 5–14, Jan. 2020.

[13] Guia, Joseph M. De, M. Devaraj and C. Leung. "DeepGx: Deep Learning Using Gene Expression for Cancer Classification." 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019

[14] Kuijjer, Marieke Lydia; Paulson, Joseph Nathaniel; Salzman, Peter; Ding, Wei; Quackenbush, John. "Cancer subtype identification using somatic mutation data." British Journal of Cancer, 2018 May 16. doi: 10.1038/s41416-018-0109-7

[15] Yang, Jungho; Min, Kyueng-Whan; Kim, Dong-Hoon; Son, Byoung Kwan; Moon, Kyoung Min; Wi, Young Chan; Bang, Seong Sik; Oh, Young Ha; Do, Sung-Im; Chae, Seoung Wan; Oh, Sukjoong; Kim, Young Hwan; Kwon, Mi Jung; Ahmad, Aamir (2018). "High TNFRSF12A level associated with MMP-9 overexpression is linked to poor prognosis in breast cancer: Gene set enrichment analysis and validation in large-scale cohorts". PLOS ONE, 13(8), e0202113–.2018

[16] McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. DePristo. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next generation DNA sequencing data." Genome research 209 (2010): 1297-303

[17] Manish, S & Bhagat, Vandana & Pramila, RM. (2021). "Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms". 2021 2nd International Conference of EmergingTechnologydoi:10.1109/INCET51464.2021.9456424.