# FEATURE EXTRACTION AND SELECTION USING GENE EXPRESSION FOR BREAST CANCER DETECTION

**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 leading diseases of death in women. It induces by a genetic mutation in breast cancer cells. Genetic testing has become popular to detect the mutation in genes but test cost is relatively expensive for several patients in developing countries like India. Genetic test takes between 2 and 4 weeks to decide the cancer. The time duration suffers the prognosis of genes because some patients have high rate of cancerous cell growth. We used a machine learning systems strategy that includes the extensive search for the discovery of the most optimal learning model, including feature selection algorithms, a feature extraction algorithm, and classifiers for diagnosing breast cancer. Hence, this study aims to obtain a high-importance transcript the profile linked with classification procedures that can facilitate the early detection of breast cancer. In the research work, a cost and time efficient method is proposed to predict the gene expression level on the basis of clinical outcomes of the patient by using machine learning techniques. An improved K Nearest Neighbour (KNN) with hyper parameter gene selection technique is proposed to find the most significant genes related to breast cancer afterward explained variance statistical analysis is applied to extract the genes contain high variance. The proposed method predicts the expression of significant genes with reduced Root Mean Square Error and acceptable adjusted R-square value. As per the study, analysis of these selected genes is beneficial to diagnose the breast cancer at prior stage in reduced cost and time.

**Keywords:** Breast cancer, Prediction, Profiling, Feature selection, gene expression analysis, gene selection, K Nearest Neighbour, Machine learning.

## 1. Introduction

Worldwide, breast cancer is the most common cancer in women, including almost one-third of all females' malignancies. The risk of developing breast cancer is a multi-step process involving multiple cell types, and its prevention remains challenging worldwide [1]. Many risk factors such as aging, estrogen, family history, and gene mutations can increase the feasibility of developing breast cancer [2]. Detection of breast cancer may be hard at the beginning of the disease due to the absence of symptoms; after some clinical tests, an accurate diagnosis should be able to differentiate the benign and malignant tumours'. Hence, an early breast cancer diagnosis is one of the best strategies to prevent this disease. In some developed countries, breast cancer patients have a 5 years relative survival rate above 80% due to early prevention [3]. Although some improvements have been achieved in the treatment methods in recent years, late diagnosis and treatment resistance [4] are serious problems that lead to poor prognoses for some patients [5].

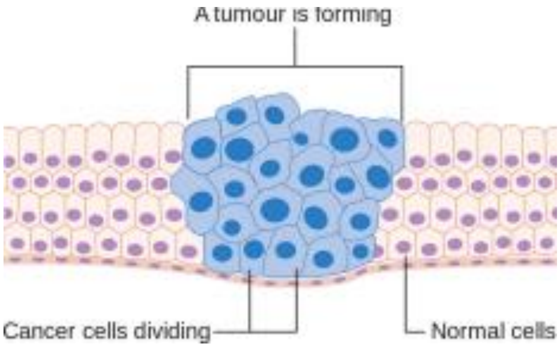


Fig 1: Cancer cells keep on forming the tumor

Different factors are involved in promoting breast cancer; most implied changes in the expression of certain genes, such as microRNAs. MicroRNAs have the ability to control signalling pathways, hence affecting tumorigenesis and various aspects of cancer progression [6]. Evaluation of the expression profiles of genes and microRNAs can be applied as valuable clues in discovering new biomarkers which are more effective for early diagnosis of breast cancer and therapeutic strategies [7]. Machine Learning (ML) procedures can be considered as another strategy for discovering valuable information in large data [8].

ML is a type of artificial intelligence (AI) that develops to simulate human intelligence by learning from data and ongoing experience. This approach needs the integration of numerous data sets of biological information, allowing the design of a statistical model that estimates the unknown parameters [9]. Recently, ML has achieved considerable success in medicine, where it has been effective in the multi-pathology classification task. In breast cancer cases, applying ML concentrates on finding specific changes in gene expression that allows earlier diagnosis of the disease [10]. Tus, an extensive data set of microRNAs can be employed in breast cancer patients as potential biomarkers and utilized them in ML algorithms. Thereby, a model can be created to detect each disease [11]. It is possible to predict the pathology more accurately and differentiate between patients with- and without breast cancer based on their information. In the present study, we used a hybrid machine learning systems (HMLS) strategy that includes the extensive search for the discovery of the most optimal HMLSs including feature selection algorithms, a feature extraction algorithm, and classifiers for diagnosing breast cancer [12]. Hence, several experiments were conducted to select the best biomarkers with the highest ranking. Therefore, this study aims to obtain a high-importance biomarker linked with classification procedures that can facilitate the early detection of breast cancer [13]. To get the best combination of feature selection and classification procedures, extensive comparative analyses were performed using performance metrics such as balanced accuracy, and receiver operator characteristic curve (AUC) statistics.

Breast cancer is a genetic disease in which cells in the breast multiply uncontrollably and become abnormal to generate a tumor. It develops as a result of genetic damage or change (mutation) in cells functioning [14]. As per the study [15], USA, China and India collectively account almost one third of global breast cancer cases whereas India has high mortality rate and low incidence rate in comparison to China and USA [16] as shown in Figure 2. In 2017, India had the highest mortality rate globally in breast cancer. In 2019, 268,600 new cases are estimated of invasive breast cancer among women and around 41,760 women died from breast cancer in 2019[17]. The major reasons of increased mortality in India are the diagnosis of cancer at last stage, inadequate screening, high-cost of screening and lack of required prevention facilities. Genetic test is an important tool to diagnose the cancer. It is basically a DNA sequencing test that compares the sequence of DNA in normal cells with cancerous cells[18]. A genetic test predicts the prognosis of genes precisely but it is expensive and time taking process to reach the final resultin developing countrieslike India[19].

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed ( Group -I) Journal Volume 11, Iss 10, Oct 2022

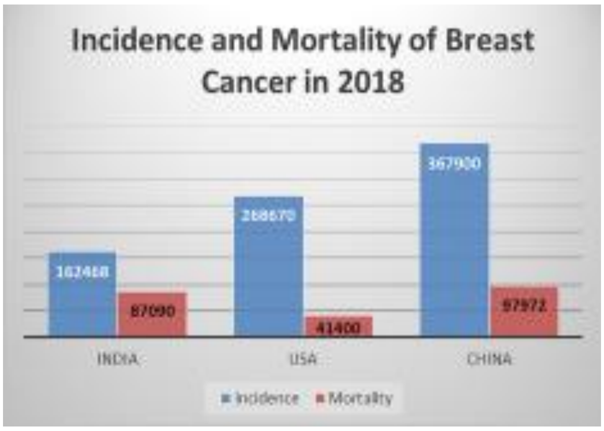


Fig 2: Breast cancer statistics

According to Rajiv Sarin, The cost of each reliable genetic test in India is expensive for several families. At present, no government hospital is providing genetic test for cancer, patient has to bear all the expense[20]. Commercial labs give the test report between 2 and 4 weeks but research canters take minimum 4 weeks or more for final report [21]. Normal cells function properly and repair themselves but cancer cells are dented cells, they do not repair themselves and assembled with the boundary of tumour [22] consequently the duration of final gene report can affect the neighbour tissues. So, a novel test is required that can provide genetic test report in reduced cost and time as time is an imperative factor in making decisions for breast cancer.

In current clinical perspective, biomarkers are involved to diagnose the cancer at the cellular level [23]. In clinical test, tissues of the suspected body parts are examined to provide evidence of the disease. It is based on certain biomarkers such as blood presser, dimension of tumour, Progesterone Receptor and Estrogen Receptor status of tumour [24]. Recent research has revealed that these biomarkers help in the prognosis and diagnosis of cancers [25]. The important part of clinical test is the understanding of alteration occurring in the cancer cells at cellular level[26].

In current state, these genomic and clinical test reports are available in the form of datasets The clinical biomarkers have been generated from gene expression data, for intense 70-gene, and 76-gene [27] signatures, and clinical data for example Nottingham Prognostic Index (NPI) and Adjuvant Online! Tools [28]. Many researchers have aimed at training model by combining both the data types such as clinical data and gene expression. From the past decade, researchers are applying Machine Learning to diagnose the cancer [29]. Machine learning is a field of Computer Science. It takes decision on the basis of past experiences using statistical and computational techniques [30].

In the proposed model, Machine Learning based SVM-RFE_MI technique is proposed for gene selection. It is an improved hybrid model of SVM-RFE and Mutual Information techniques that provides the significant genes associated with breast cancer. Along with this, clinical evidence of breast cancer patient is used to predict the expression level of selected genes by applying Least Absolute Shrinkage Selector Operator (LASSO) regression technique. The goal of lasso regression is to attain the subset of predictors that reduces the prediction error for a measurable response variable. The research work is described in four sections, next section describes the related work, second section deals with the methods and materials used in the work, third section deals with the experimental results and fourth section describes the conclusion of the proposed work followed with reference section.

The description of genes is as follows.

Gene 1- AUREKA: It is a Kinase, that is important for cell division. Its main function is regulating the mitosis specifically chromosomal segregation. The mutation in AURKA kinase leads to failure of cell division and harm the progression of cells.

Gene 2- GSTM1: It is antioxidant. It converts free radicals into molecules. Mutations of this gene lead to unstable the free radicals. These free radicals damage the cells and converts into cancerous cells.

Gene 3- IGFBP5: It is an Insulin-like protein-binding growth factor that plays a vital role in cell growth, differentiation and apoptosis. Its key role is cell regulation and breast tissue development. IGFBBPS mutation could lead to differentiation of breast tissue and development of cancer.

Gene 4- BCL2: It is Apoptosis Regulator. Cancer cell depends on this gene to survive. BCL2 needs to remove so that cells can undergo programmed cell death.

Gene 5- VEGF: It is an endothelial vascular growth factor. It induces mitogenesis, survival of endothelial cells, stromal degradation and vascular permeability. Over expression of VEGF lead to tumor development and neovascularization.

Gene 6- RRM2: It is ribonucleotide reductase regulatory subunit, which catalyzes the development of ribonucleotide deoxy ribo nucleotides. RRM2 regulate the cell cycle by synthesis and degradation of DNA and RNA. Inhibitions of this enzyme in cancer patients considerably reduce cell cycle gene expression.

## 2. Literature Review

Masses are the primary indications of breast cancer in mammograms, and it is important to classify them as benign or malignant. Benign and malignant masses differ in geometry and texture characteristics. However, not every geometry and texture feature that is extracted contributes to the improvement of classification accuracy; thus, to select the best features from a set is important. In this paper, Liu et al. [1] examined the feature selection methods for mass classification. We integrate a support vector machine (SVM)-based recursive feature elimination (SVM-RFE) procedure with a normalized mutual information feature selection (NMIFS) to avoid their singular disadvantages (the redundancy in the selected features of the SVM-RFE and the un optimized classifier for the NMIFS) while retaining their advantages, and we propose a new feature selection method, which is called the SVM-RFE with an NMIFS filter (SRN). In addition to feature selection, we also study the initialization of mass segmentation.

Heterogeneity in cancer can affect response to therapy and patient prognosis. Histologic measures have classically been used to measure heterogeneity, although a reliable non-invasive measurement is needed both to establish baseline risk of recurrence and monitor response to treatment. Here, Mahrooghy et al. [2] proposed using spatiotemporal wavelet kinetic features from dynamic contrast-enhanced magnetic resonance imaging to quantify intra tumour heterogeneity in breast cancer. Tumour pixels are first partitioned into homogeneous sub regions using pharmacokinetic measures. Heterogeneity wavelet kinetic (HetWave) features are then extracted from these partitions to obtain spatiotemporal patterns of the wavelet coefficients and the contrast agent uptake. The HetWave features are evaluated in terms of their prognostic value using a logistic regression classifier with genetic algorithm wrapper-based feature selection to classify breast cancer recurrence risk as determined by a validated gene expression assay.

Due to the vital role of the aberrant DNA methylation during the disease development such as cancer, the comprehension of its mechanism had become essential in the recent years for early detection and diagnosis. With the advent of the high-throughput technologies, there are still several challenges to achieve the classification process using the DNA methylation data. The high-dimensionality and high-noisiness of the DNA methylation data may lead to the degradation of the prediction accuracy. Thus, it becomes increasingly important in a wide range to employ robust computational tools such as feature selection and extraction methods to extract the informative features amongst thousands of them, and hence improving cancer prediction. By using the DNA methylation degree in promoters and probes regions, Raweh et al. [3] aimed at predicting cancer with a hybridized approach based on the feature selection and feature extraction techniques. The suggested approach exploits a filter feature selection method called (F-score) to overcome the high-dimensionality problem of the DNA methylation data, and proposes an extraction model which employs the peaks of the mean methylation density, the fast Fourier transform algorithm, and the symmetry between the methylation density of a sample and the mean methylation density of both sample types normal and cancer as novel feature extraction methods, in order to accurate cancer classification and reduce training time.

In shear wave absolute vibro-elastography (S-WAVE), a steady-state multi-frequency external mechanical excitation is applied to tissue, while a time-series of ultrasound radio-frequency (RF) data are acquired. Shao et al. [4] objective is to determine the potential of S-WAVE to classify breast tissue lesions as malignant or benign. We present a new processing pipeline for feature-based classification of breast cancer using S-WAVE data, and we evaluate it on a new data set collected from 40 patients. Novel bi-spectral and Wigner spectrum features are computed directly from the RF time series and are combined with textural and spectral features from B-mode and elasticity images. The Random Forest permutation importance ranking and the Quadratic Mutual Information methods are used to reduce the number of features from 377 to 20. Support Vector Machines and Random Forest classifiers are used with leave-one-patient-out and Monte Carlo cross-validations. Classification results obtained for different feature sets are presented.

Early detection and diagnosis of breast cancer are crucial to improve the survival rates of patients. Hence, pathologists and radiologists need a computer-aided diagnosis system to assist their clinical diagnoses effectively and efficiently. However, most breast cancer recognition models are faced with the sample scarcity problem, which results in serious over fitting and lowers recognition performance. To alleviate the sample scarcity problem, a simple, effective model called “refinement, correlation, adaptive” (RCA) for breast cancer recognition is proposed from the perspective of fine-grained feature selection. An innovative multi view efficient range-based gene selection algorithm is proposed by Li et al. [5] to complete the first-layer feature “refinement,” which contributes to suppressing the noisy information in the original feature space. Then, more-discriminant but low-dimensional information among heterogeneous features is mined through the second-layer cross-modal “correlation” mining. Feature dimensions are reduced to a reasonable value that fits the sample size well and alleviates the over fitting problem. Finally, the last-layer decision-tree-guided “adaptive” feature selection is completed using the gradient boosting decision tree algorithm. The RCA model was validated on two well-known datasets.

Breast cancer is one of the most common cancers diagnosed in women. For preventive diagnosis, feature selection is an essential step to construct the breast cancer classifier. The features of a real breast cancer dataset are usually composed of discrete and continuous ones. Also, the Area Under the Curve (AUC) of the receiver operating characteristic receives more attention in such a medical field. The existing research work is insufficient to take into account both the hybrid trait of the features and the specific classification objective. Wuniri et al. [6] proposed a wrapper method, i.e., a integrated framework in which Bayesian classifiers are embedded for the feature selection of breast cancer datasets. To deal with both the discrete features and the continuous features, we adopt the naive approach for the discrete features but the kernel probability density estimation for the continuous ones, respectively, which leads to feature-type-aware hybrid Bayesian classifiers. All the classifiers are fed with different feature subsets and evaluated by their AUC metrics as the fitness indexes. Thus, with the genetic algorithm, we can obtain a near optimal feature subset, which yields a good AUC metric with its corresponding classifiers. Moreover, the one-class F-score is used to help enhance the convergence of the algorithm.

Breast cancer is a neoplastic disease which seriously threatens women's health. It is regard as the most common cause of cancer death in women. Accurate detection and effective treatment are of vital significance to lower the death rate of breast cancer. In recent years, machine learning technique has been considered to be an effective method for accurate diagnosis of various diseases, among which Random Forest (RF) has been widely applied. However, decision trees with poor classification performance and high similarity may be generated during the training process, which affects the overall classification performance of the model. In this paper, a Hierarchical Clustering Random Forest (HCRF) model is developed by Huang et al. [7]. By measuring the similarity among all the decision trees, the hierarchical clustering technique is used to carry out clustering analysis on decision trees. The representative trees are selected from divided clusters to construct the hierarchical clustering random forest with low similarity and high accuracy. In addition, we use Variable Importance Measure (VIM) method to optimize the selected feature number for the breast cancer prediction. Wisconsin Diagnosis Breast Cancer (WDBC) database and Wisconsin Breast Cancer (WBC) database from the UCI (University of California Irvine) Machine Learning repository are employed in this study. The performance of the proposed method is evaluated by utilizing accuracy, precision, sensitivity, specificity and AUC Under ROC Curve).

Histopathological image analysis is an important technique for early diagnosis and detection of breast cancer in clinical practice. However, it has limited efficiency and thus the detection of breast cancer is still an open issue in medical image analysis. To improve the early diagnostic accuracy of breast cancer and reduce the workload of doctors, we devise a classification framework based on histology images by combining deep learning with machine learning methodologies in this paper. Specifically, Wang et al. [8] devised a multi-network feature extraction model by using pre-trained deep convolution neural networks (DCNNs), develop an effective feature dimension reduction method and train an ensemble support vector machine (E-SVM). First, we pre-process the histological images via scale transformation and color enhancement methods. Second, the multi-network features are extracted by using four pre-trained DCNNs (e.g., DenseNet-121, ResNet-50, multi-level InceptionV3, and multi-level VGG-16). Third, a feature selection method via dual-network orthogonal low-rank learning (DOLL) is further developed for performance boosting and over fitting alleviation. Finally, an E-SVM is trained via fused features and voting strategy to perform the classification task, which classifies the images into four classes.

## 3. Dataset Description

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

## 4. Proposed Method

ML methods are tools utilized to create and evaluate algorithms that facilitate prediction and classification. ML is based on four steps: data collection, picking a model, training the model, and testing the model. In this study, three groups of algorithms were employed: feature selection, feature extraction, and classification algorithms. We employed feature extraction to convert high-dimensional data to fewer dimensions; thus, the risk of over fitting was reduced. Dimensionality reduction procedures use no label for feature extraction; hence, they only rely on patterns between input features. Consistent with previous studies [13] Principal Component Analysis (PCA) was outperformed other feature extraction algorithms. PCA is a dimensionality reduction procedure that generates new specified features, but not a feature selection procedure. PCA transforms features, but feature selection procedures choose features without transforming them. Hence, as the feature extractor, PCA was implemented in this study.

ML procedures have difficulty in dealing with a large number of input features. Hence, to support the process of applying ML effectively in real-world scenarios, data pre-processing is an essential task. Feature selection is one of the most frequent procedures in data pre-processing [18]. It has become a vital element of the ML process to obtain the relevant feature or feature subsets in the literature to achieve their classification objectives. However, besides the advantages of feature selection procedures to search for a subset of relevant features, they are used to avoid over fitting and provide faster and more cost-effective models.

The cross-combination approach was employed to compare the performance of feature selection, extraction, and classification procedures. Therefore, each feature selection and the extraction procedure were combined with all the nine classification procedures. Finally, we got 65 combinations of ML strategies. The Work flow of the proposed work has shown in below Figure 3.

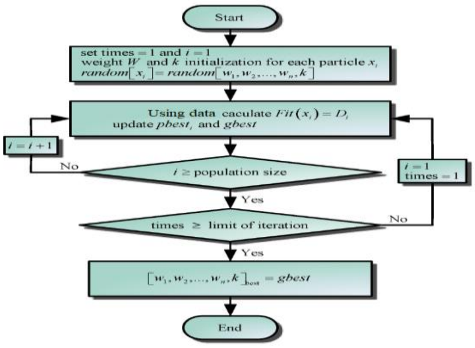

Start
set times = 1 and $i = 1$
weight $W$ and $k$ initialization for each particle $x_i$
$random[x_i] = random[w_1, w_2, \ldots, w_n, k]$
Using data caculate $Fit(x_i) = D_i$
update $pbest_i$ and $gbest$
$i = i + 1$
No
$i \geq$ population size
$i = 1$
times = 1
Yes
times $\geq$ limit of iteration
No
Yes
$[w_1, w_2, \ldots, w_n, k]_{best} = gbest$
End

Fig 3: Work flow of the proposed work.

An improved K Nearest Neighbour (KNN) gene selection technique is proposed to find the most significant genes related to breast cancer afterward explained variance statistical analysis is applied to extract the genes contain high variance. K-Nearest Neighbors (KNN)

KNN is a Machine Learning Algorithm. In this Classifier each new Document is compared to fundamental documents. This Algorithm checks how a document is classified by looking at only training data that are equal to it. KNN assumes that to divide the documents like a points in the Euclidean area. The distance among the two points of any plane with the p(x,y) and q(a,b) calculated as :

$$d = \sqrt{(x-a)^2 + (y-b)^2}$$

**Dataset Loading:** The required dataset, for this thesis 20 Newsgroups dataset, is loaded during execution either directly from internet or from local system on which it is present previously. Case folding is applied to convert all characters into same case, lower case, in order to avoid duplication of words.

**Removal of Header/Footers/Quotes:** All the documents of the dataset across all categories contain headers/footers/quotes such for example From, Subject, Signature Line etc. These segments need to be eliminated from the actual content in order to avoid over fitting and generate a more generalized classifier model for better classification.

**Tokenization:** In this each and every document is treated as a string, and then partitioned into tokens containing characters mentioned in condition of regular expression. Tokenization includes separating sequence of strings into phrases, keyword, phrases, words, symbols called tokens. Punctuation marks are not considered in tokenization.

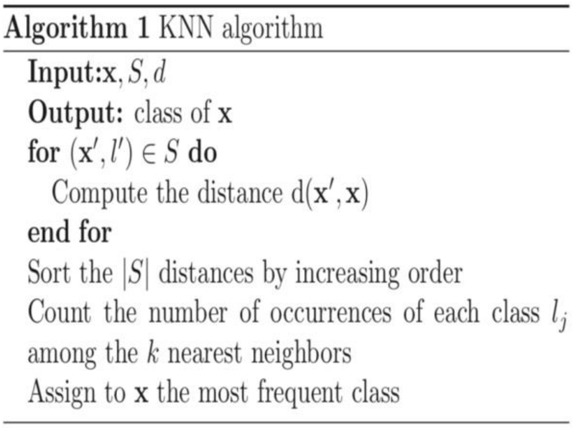
Feature Representation and Extraction Vector space model is the most common method used for document representation. Here each document is represented as a vector d and each dimension in the vector corresponds to a distinct term, called as features, in the term space of the document collection. This representation is expressed as: d = (w1; w2; ...; wn) where wi denotes weight of term i in document d. To compute these term weights several methods are formulated. For proper term weighting and feature extraction, we need a method which considers rare terms as similarly as frequent terms, multiple appearances of a term in a document to be more important than single appearances, and is not biased towards long documents.

One of the widely used weighting methods taking these properties into account is the term frequency-inverse document frequency (TF-IDF) weighting. Calculation of 'df' resembling 'document frequency', 'tf' resembling 'term frequency' and length normalization present in this formula considers the above stated properties simultaneously. Thus, in our work we apply TF-IDF method whose formula is given below: $w_{ij} = tf_{ij} \log(N/df_i)$ (3) Here, in a document j the weight of a term is i is wij, frequency of a term i is $tf_{ij}$ in a document j and $df_{ij}$ is the number of documents in which a term i occurs in the whole document collection. N is the whole number of documents.

In tf-idf weighting method, if a term often occurs in a document, it is more discriminative whereas if it appears in most of the documents, then it is less discriminative for the content. This constructed vector space model improves the accuracy, efficiency and scalability of the classification model. Feature Selection In the text documents, the high dimensionality of features or terms reduces the accuracy of classification due to irrelevant features. Feature selection, other than feature extraction, is one of the well-known methods of reducing the dimensionality by removing non-informative words. These irrelevant and misleading words are found by ranking all features according to their importance estimated by a metric and then selecting ones with higher values.

The top words extracted out are then used to classify the documents. Hence, to select features from documents Feature selection techniques are used. It aims at reducing time and improving efficiency of classifiers by removing noise features. Two main policies are used in feature selection viz. global and local policies while in the second one, a different set of features is selected from each class. In global policy, a single set of features is selected from all classes which provide a global view of entire dataset by extracting a single global score from local scores. Thus, it tends to penalize the infrequent classes in highly skewed datasets. In local policy, a different set of features is selected from each class which tends to give equal weightage to each one of them and thus, it optimizes the performance of classification on frequent and infrequent classes. The steps involved in KNN are discussed.

1. Calculate "$d(x, x_i)$" i =1, 2, ….., n; where d denotes the Euclidean distance between the dataset points.
2. Arrange the calculated n Euclidean distances in non-decreasing order.
3. Let k be a +ve integer, take the first k distances from this sorted list.
4. Find those k-points corresponding to these K-distances.
5. Assign weights to the nearest k-distance data points.
6. Let $k_i$ denotes the number of points belonging to the ith class among k points i.e. $k \geq 0$
7. If $k_i > k_j$ $\forall$ $i \neq j$ then put x in class i.



**Algorithm 1** KNN algorithm

```
Input: $\mathbf{x}, S, d$
Output: class of $\mathbf{x}$
for $(\mathbf{x}', l') \in S$ do
    Compute the distance $d(\mathbf{x}', \mathbf{x})$
end for
Sort the $|S|$ distances by increasing order
Count the number of occurrences of each class $l_j$ among the $k$ nearest neighbors
Assign to $\mathbf{x}$ the most frequent class
```

## 5. Results

Breast cancer is the second largest cancer in the world, the incidence of breast cancer continues to rise worldwide, and women's health is seriously threatened. Therefore, it is very important to explore the characteristic changes of breast cancer from the gene level, including the screening of differentially expressed genes and the identification of diagnostic markers. We determined that three oncogenes, PD-L2, ETV5, and MTOR and 113 long intergenic non-coding RNAs (lincRNAs) were constantly up-regulated, whereas two oncogenes, BCR and GTF2I, one tumour suppression gene MEN1, and 30 lincRNAs were constantly down-regulated. Up-regulated genes were enriched in "focal adhesion" and "PI3K-Akt signalling" pathways, etc., and down-regulated genes were significantly enriched in "metabolic pathways" and "viral myocarditis".

This work is experimented on WBCD (Wisconsin breast cancer database) [10, 3] obtained by the University of Wisconsin. The database comprises of breast cancer data taken from the human breast tissue. There are totally 699 cases, of which 458 cases are benign tumours' and 241 are malignant cases. In the WBCD dataset, there are around 16 instances that are having missing values. These records are deleted and we are left with 683 clinical cases. We have applied the k-Nearest Neighbor algorithm investigating the several variants of distance metrics, different K values and different classification rules. This algorithm does not require learning phase. In order to evaluate the performance of the method, the dataset is partitioned into Training Set and Testing Set. Training set with class labels are used to train the model.

Eight up-regulated genes exhibited doubled or higher expression and the expression of three down-regulated genes was halved or lowered and correlated with long-term survival. Gene expression signatures were initially developed to take into account tumour biology for adjuvant chemotherapy decision and have become a standard option in hormone receptors-positive/HER2-negative early breast cancer. This prospective review highlights the unsolved issues regarding targeted populations, delineates the best clinical indications and addresses questions that ongoing and future trials will have to meet. Apart from adjuvant chemotherapy indications, we review their potential interest to tailor neoadjuvant systemic treatments, adjuvant radiation therapy, extended adjuvant hormone therapy and CDK4/6 inhibitor adjuvant treatment.

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. The feature extraction time levels of proposed and existing models are shown in Figure 4.

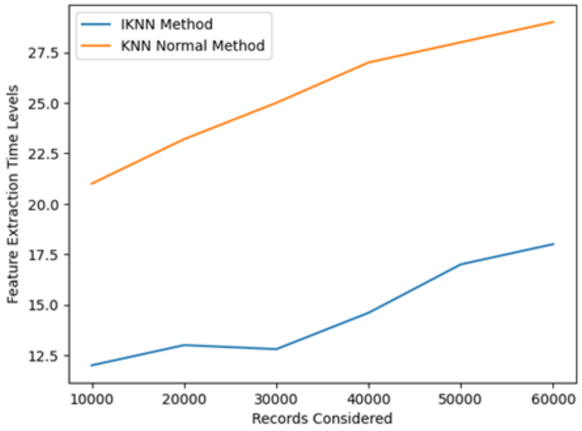


Fig 4: Feature Extraction Time Levels

The nearest points from the dataset are calculated using the Euclidean distance and the comparisons is represented in Figure 5.

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper* © 2012 IJFANS. All Rights Reserved, UGC CARE Listed ( Group -I) Journal Volume 11, Iss 10, Oct 2022

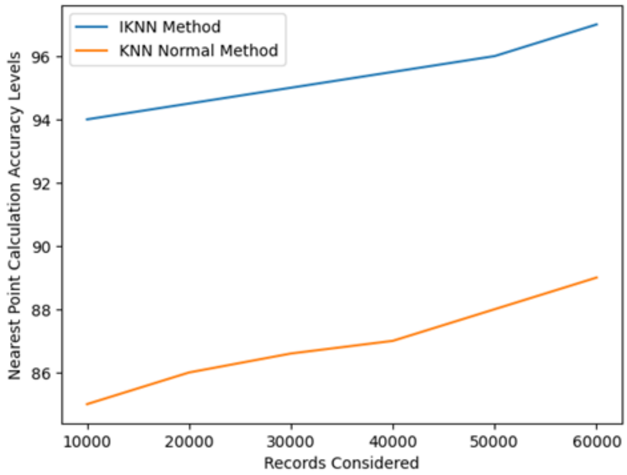


Fig 5: Nearest Point Calculation Accuracy Levels

TO each nearest point, weights are allocated so that features with highest weights are considered for training and the weight allocation is performed that is represented in Figure 6.

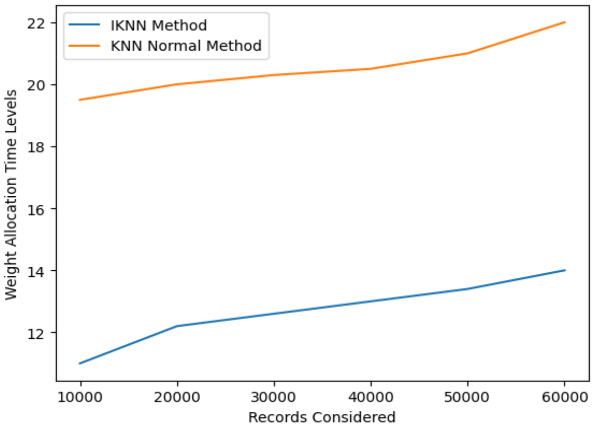


Fig 6: Weight Allocation Time Levels

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. The feature selection time levels are depicted in Figure 7.

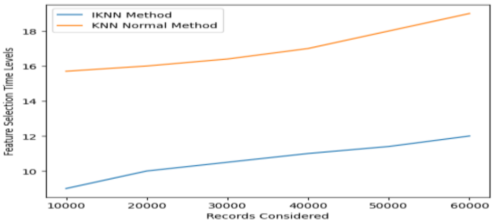


Fig 7: Feature Selection Time Levels

The proposed model performance metrics are compared with the traditional model and the results are shown in Figure 8.

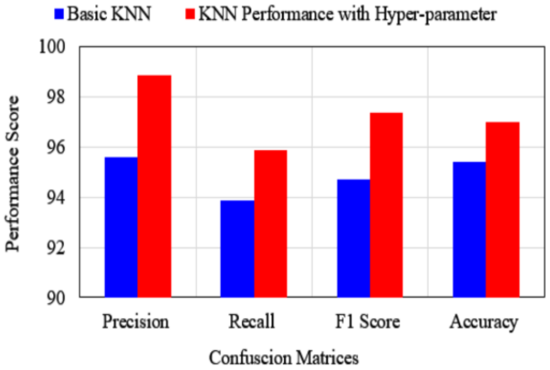


Fig 8: Evaluation Metrics

## 6. Conclusion

Breast cancer is frequent occurring cancer among women. It occurs due to the mutation in genes, therefore genetic test is preferred to identify the gene mutation. Genetic test report is accurate to detect the tumour stage but the test is expensive and time consuming in developing countries like India. In this research work, a novel test method is generated to predict the gene expression in reduced cost and time. This test method generates the expression of most significant genes from clinical outcome and provides the prognosis stage of cancer. The results show that adjusted RMSE and R-Squared values lies within standard acceptable range. Selecting the value of K in K-nearest neighbour is the most critical problem. A small value of K means that noise will have a higher influence on the result i.e., the probability of over fitting is very high. A large value of K makes it computationally expensive and defeats the basic idea behind KNN. A simple approach to select k is k = n^(1/2). To optimize the results, we can use Cross Validation. Using the cross-validation technique, we can test KNN algorithm with different values of K. The model which gives good accuracy can be considered to be an optimal choice. It represents that the test method has good prediction accuracy. This test method will provide outcome immediately after final clinical report with no cost. It is useful for all the patients suffering from breast cancer. The proposed test method leads to reduce the mortality by identifying the cancer at earliest phase. At the end we got the best future subsets for doing classification and clustering for each gene expression dataset.

## References

[1] 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.

[2] M. Mahrooghy et al., "Pharmacokinetic Tumor Heterogeneity as a Prognostic Biomarker for Classifying Breast Cancer Recurrence Risk," in IEEE Transactions on Biomedical Engineering, vol. 62, no. 6, pp. 1585-1594, June 2015, doi: 10.1109/TBME.2015.2395812.

[3] A. A. Raweh, M. Nassef and A. Badr, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation," in IEEE Access, vol. 6, pp. 15212-15223, 2018, doi: 10.1109/ACCESS.2018.2812734.

[4] Y. Shao et al., "Breast Cancer Detection Using Multimodal Time Series Features From Ultrasound Shear Wave Absolute Vibro-Elastography," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 2, pp. 704-714, Feb. 2022, doi:10.1109/JBHI.2021. 3103676.

[5] G. Li et al., "Effective Breast Cancer Recognition Based on Fine-Grained Feature Selection," in IEEE Access, vol. 8, pp. 227538-227555, 2020, doi:10.1109/ACCESS.2020. 3046309.

[6] Q. Wuniri, W. Huangfu, Y. Liu, X. Lin, L. Liu and Z. Yu, "A Generic-Driven Wrapper Embedded With Feature-Type-Aware Hybrid Bayesian Classifier for Breast Cancer Classification," in IEEE Access, vol. 7, pp. 119931-119942, 2019, doi:10.1109/ ACCESS.2019.2932505.

[7] Y. Wang et al., "Breast Cancer Image Classification via Multi-Network Features and Dual-Network Orthogonal Low-Rank Learning," in IEEE Access, vol. 8, pp. 27779-27792, 2020, doi: 10.1109/ACCESS.2020.2964276.

[8] P. Liu and S. Fei, "Two-Stage Prediction of Comorbid Cancer Patient Survivability Based on Improved Infinite Feature Selection," in IEEE Access, vol. 8, pp. 169559-169567, 2020, doi: 10.1109/ACCESS.2020.3016998.

[9] X. Tang, L. Cai, Y. Meng, C. Gu, J. Yang and J. Yang, "A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis With Transcriptome and Functional Proteomic," in IEEE Access, vol. 9, pp. 51659-51668, 2021, doi: 10.1109/ACCESS.2021.3070428.

[10] S. Liu et al., "Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall," in IEEE Access, vol. 9, pp. 24433-24445, 2021, doi: 10.1109/ ACCESS.2021.3054823.

[11] A. U. Haq, D. Zhang, H. Peng and S. U. Rahman, "Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection," in IEEE Access, vol. 7, pp. 151482-151492, 2019, doi: 10.1109/ACCESS.2019.2947701.

[12] M. Qiao et al., "Breast Tumor Classification Based on MRI-US Images by Disentangling Modality Features," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 7, pp. 3059-3067, July 2022, doi: 10.1109/JBHI.2022.3140236.

[13] A. F. Al-Juniad, T. S. Qaid, M. Y. H. Al-Shamri, M. H. A. Ahmed and A. A. Raweh, "Vertical and Horizontal DNA Differential Methylation Analysis for Predicting Breast Cancer," in IEEE Access, vol. 6, pp. 53533-53545, 2018, doi: 10.1109/ACCESS.2018.2871027.

[14] B. Fu, P. Liu, J. Lin, L. Deng, K. Hu and H. Zheng, "Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 7, pp. 2053-2064, July 2019, doi: 10.1109/TBME.2018.2882867.

[15] S. B. Sakri, N. B. Abdul Rashid and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," in IEEE Access, vol. 6, pp. 29637-29647, 2018, doi: 10.1109/ACCESS.2018.2843443.

[16] N. E. M. Khalifa, M. H. N. Taha, D. Ezzat Ali, A. Slowik and A. E. Hassanien, "Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach," in IEEE Access, vol. 8, pp. 22874-22883, 2020, doi: 10.1109/ACCESS.2020.2970210.

[17] D. B. Fogel, E. C. Wasson, E. M. Boughton, V. W. Porto and P. J. Angeline, "Linear and neural models for classifying breast masses," in IEEE Transactions on Medical Imaging, vol. 17, no. 3, pp. 485-488, June 1998, doi: 10.1109/42.712139.

[18] S. B. Ginsburg, G. Lee, S. Ali and A. Madabhushi, "Feature Importance in Nonlinear Embeddings (FINE): Applications in Digital Pathology," in IEEE Transactions on Medical Imaging, vol. 35, no. 1, pp. 76-88, Jan. 2016, doi: 10.1109/TMI.2015.2456188.

[19] M. W. A. El-Soud, I. Zyout, K. M. Hosny and M. M. Eltoukhy, "Fusion of Orthogonal Moment Features for Mammographic Mass Detection and Diagnosis," in IEEE Access, vol. 8, pp. 129911-129923, 2020, doi: 10.1109/ACCESS.2020.3008038.

[20] B. Yousefi, H. M. Sharifipour and X. P. V. Maldague, "A Diagnostic Biomarker for Breast Cancer Screening via Hilbert Embedded Deep Low-Rank Matrix Approximation," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-9, 2021, Art no. 4504809, doi: 10.1109/TIM.2021.3085956.

[21] 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.

[22] Z. Zhou, S. Li, G. Qin, M. Folkert, S. Jiang and J. Wang, "Multi-Objective-Based Radiomic Feature Selection for Lesion Malignancy Classification," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 1, pp. 194-204, Jan. 2020, doi: 10.1109/JBHI.2019.2902298.

[23] M. H. Waseem et al., "On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction," in IEEE Access, vol. 7, pp. 141072-141082, 2019, doi: 10.1109/ACCESS.2019.2944295.

[24] J. Yao, J. Chen and C. Chow, "Breast Tumor Analysis in Dynamic Contrast Enhanced MRI Using Texture Features and Wavelet Transform," in IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 1, pp. 94-100, Feb. 2009, doi: 10.1109/JSTSP.2008.2011110.

[25] B. Zeimarani, M. G. F. Costa, N. Z. Nurani, S. R. Bianco, W. C. De Albuquerque Pereira and C. F. F. C. Filho, "Breast Lesion Classification in Ultrasound Images Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 133349-133359, 2020, doi: 10.1109/ACCESS.2020.3010863.

**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,Thirupathi. 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 andInternational Journals. His research areas of interest are MachineLearning, Data Mining, Data warehousing and Data Mining.

**Dr. R.J.Rama Sre**e, 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