HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

A Novel Computer Assisted Genomic Test method to Detect Breast Cancer in Reduced Cost and Time using Ensemble Technique

Madhuri Gupta[1], Bharat Gupta[2], Ishan Budhiraja[3], Deepak Garg[4], Ketan Kotecha[5], and Celestine Iwendi[6]

Abstract

Breast cancer is the leading cause of death among women around the world. It is a primary malignancy for which genetic markers have revealed the ability for clinical decision making. It is a genetic disease that generates due to gene mutations, but the cost of a genetic test is relatively high for a number of patients in developing nations like India. The results of a genetic test can take a few weeks to determine cancer. This time duration influences the prognosis of genes since certain patients suffer from a high rate of malignant cell proliferation. Therefore, a computer-assisted genetic test method (CAGT) is proposed to detect breast cancer. This test method will predict the gene expressions and convert these expressions in the state of mutation (under-expression (-1), transition (0) overexpression (1)) and afterwards perform the classification to get the benign and malignant class in reduced time and cost. In the research work, machine learning techniques are applied to identify the most responsive genes of breast cancer on the premises of the clinical report of a patient and generated a CAGT. In the research work, the hard voting ensemble approach is applied to detect breast cancer on the basis of most responsive genes by CAGT which leads to improving 3.5% accuracy in cancer classification.

Keywords

Electronic health record; Breast Cancer, Machine Learning, Ensemble modeling, Genomics

1. Introduction

Cancer is a genetic disease. It is categorized by the uncontrolled growth of cells in the body due to mutations in genes [1]. It can develop anywhere in the body. It starts when a few clusters of cells grow uncontrollably and crowd out the neighbouring cells for resources. Breast cancer is the frequently diagnosed form of cancer. It has consistently been a leading cause of cancer fatalities for decades. Once metastasized, affected cancerous cells are capable to infiltrate any organ of the body and hamper its normal course of functioning [2]. It, therefore, becomes imperative to diagnose cancer at a primary stage.

Cancer is a genetic disease so genomic test gives the more accurate detection of cancer. Gene expression report contains genes mutation information, but the genomic test is costly and time-consuming in underdeveloped nations like India. Genetic tests are unaffordable for several families in India [3] and

***Corresponding Author:** Ketan Kotecha (head@scaai.siu.edu.in) and Ishan Budhiraja (ishan.budhiraja@bennett.edu.in)
[1,3,4]School of Computer Science Engineering & Technology (SCSET), Bennett University, Greater Noida.
[2]Department of CS&IT, Jaypee Institute of Information Technology, Noida.
[5]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Lavale, Pune, Maharashtra, India.
[6]Computing Dept. School of Creative Technologies, University of Bolton, Bolton, UK, A676 Deane Rd, Bolton BL3 5AB, United Kingdom

most of the expenses have to be borne by patients as very few government hospitals are offering genetic tests. Research centers with the infrastructure to offer genetic testing usually provide results in 4 weeks while private testing facilities generate the test report in 2 to 4 weeks. This idle time spent to wait for the reports can lead to cancer growth and increased spread throughout the body [3]. So, a test method is required that can diagnose breast cancer on the basis of gene mutation in a reduced time.

In the research work, a Computer-Assisted Gene Test (CAGT) method is generated that can predict the most responsive genes on the basis of the clinical report of patients. To generate CAGT, a Van't Veer data set is used that contains the clinical report and gene expression of the same patients. The clinical report contains the test report of different biomarkers that are involved in cancer such as follow-up time (year), diameter (mm), metastases, angioinvasion, grade, lymphocytic infiltrate, ERp, PRp and brca1 mutation. The Gene expression dataset contains the genomic information of the patient. It is generated by microarray technology [4]. In the genomic research area, microarray technology is a widely used tool for gene identification of specific therapy [5]– [7]. It is mainly applicable for cancer prediction [8], [9], drug action investigation [10] and disease diagnosis [11]. The microarray technique offers the high-dimensional data that comprises the genes expression information with various environmental conditions as a large matrix.

High-throughput technology is used to test the genes that certainly generate the gene expression data in a large amount. It leads to the incorporation of data analysis techniques, machine learning and tools to extract details from large data. Machine Learning (ML) and data analytics are key tools in breast cancer research when it comes to detection and other clinical issues. ML emphasizes the advancement of computer algorithms that process data and enable the machine to learn [16], [17]. In the research work, gene expression microarray and clinical data of the same patient are analyzed and processed for an advanced testing model.

The following point summarizes the research contribution:

- We propose an SVM-RFE_MI technique to select the most relevant genes in the dataset.
- Regression technique is tuned to predict the gene expressions of selected genes on the basis of clinical parameters.
- Classification model is trained to identify cancer using predicted gene expression. The proposed solution is performed using Apache Spark data processing engine and shows promising results of cost and time reduction.

The rest of the paper is organized as follows: Section 2 depict the related works and explain the main

Algorithms, technologies, and data processing engines are used in this framework. The proposed methodology is described in Section 3 with a detailed explanation of the process flow. Section 4 presents the experimental design of the proposed framework. Section 5 represents the experimental results and the work is concluded with Section 5.

2. Related Work

Breast Cancer is the most common disease in women that is caused by the excessive proliferation of

mutated cells (Breast State, 2017). It starts to begin from breast tissues and spread to remaining fatty tissues. Breast cancer is generally detected when patients feel the lump or during breast screening. The breast lumps can be cancerous (malignant) and non-cancerous (benign). Researchers have applied different approaches to predict cancer at an early stage to decrease the mortality due to breast cancer. During the past few decades, machine learning methods have been implemented into various complex, complicated and massive data-intensive fields, for instance, cosmology, prescription, cellular biology etc. These approaches provide precise solutions to extract the information encrypted in the data [27], [28], [49], [80], [81]. Machine Learning contains various techniques which are specifically designed to assist in cancer diagnosis. A survey of machine learning is presented on the basis of techniques, metrics and frameworks used for breast cancer detection. ML techniques involved to diagnose the cancer are represented in Table 1.

Table 1: Study of Machine learning approaches applied in breast cancer prediction (State-of-art)

S. no.	**Machine learning methods**	**Solutions**	**Outcome (%)**
1	Support Vector Machine technique [12]	Breast cancer prediction using Microarray Dataset	Accuracy: 94.5%
2	Deep learning Technique [13].	Prediction of cancer survival rate using multi-platform data analysis.	mean survival in breast cancer: 70%
3	Deep Transfer Learning using single cell [14].	Breast Cancer high-content analysis	Accuracy: 77%
4	Artificial Neural Networks (ANN) [15].	Risk estimation of breast cancer	Accuracy: 96%
5	Ensemble learning using ANN with C4.5 Rule [16].	Disease diagnosis (diabetes, hepatitis and Breast cancer)	Error rate: 2.9, 24 & 14.9 respectively.
6	Simple Logistic, RBF Network and RepTree[17].	Prediction of breast cancer Survivability	Accuracy: 74.5%
7.	Analogous Random Forest Technique [18].	Big Data Computing Analytic Using Spark Cloud Computing	Error rate is 0.2 for 500 trees.
8.	Neural Networks (NN) [19].	Limited datasets handling for medical diagnosis	Accuracy: 86.5%
9.	Artificial Neural Networks re-entered [15].	Risk estimation of breast cancer	Accuracy: 96%
10.	Support Vector Machine [20].	Breast cancer prediction and susceptibility using nucleotide polymorphisms	predictive power: 69%
11.	Semi-supervised learning technique based on graph [21].	Breast cancer survivability prediction using semi-supervised learning, SVM and ANN models.	Accuracy: 76%
12.	Graph based semi-supervised machine learning technique [22].	Generate an integrated gene network model to observe cancer recurrence.	Max Accuracy**: 80%**
13.	Semi Supervised Learning Co-Training Algorithm [23].	Breast cancer survivability prediction	76% Accuracy
14.	Support Vector Machine [24].	Breast cancer prediction using gene recurrence and signature	73% Specificity 89% Sensitivity
15.	Bayesian network [25].	Incorporation of microarray and clinical data to predict the breast cancer.	85% AUC
16.	Executed 17 support vector machine classifier model using LIBLINEAR [26]	Cancer Classification at primary sites	62% Accuracy

The study presented in table 1 enlists commonly used Machine Learning techniques for gene selection. Machine learning techniques are applicable in various data-intensive fields like cosmology, biology and prescription, etc. In the research work, Machine learning techniques are applied to process the gene expression microarray and clinical data for the advanced testing models.

Data acquisitions of gene expressions are accomplished by High-Throughput Microarray Technology (HTT). The microarray is a commonly used technology in the genomic research area. It is mainly applicable in disease prediction, gene identification for a drug action investigation, specific therapy and cancer prediction. The microarray technique provides high-dimensional data that comprises the gene expression information with different environmental conditions. In the past decades, several researchers have found out the insight of genomic datasets.

To process the high throughput microarray dataset, above mentioned technologies need the data processing engine. These engines are capable to process different sizes of data [72]. The selection of data processing engines depends on the significance and size of the dataset. A study of some data processing engines is shown below:

Data Processing Engines

In the past decade, several tools are accessible data processing such as Map Reduce, Storm, Apache Spark, Apache Flink and H2O. The study [65] of these widely used data processing engines are as follows:

Apache MapReduce and Apache Hadoop: Apache Hadoop and MapReduce both are different paradigms. MapReduce is a programming platform that processes high dimensional data by using the divide and conquer approach. Hadoop is an open-source platform that is the implementation MapReduce.

Apache Spark: Apache spark is developed for speed processing and big data analytics. Spark is an open-source in-memory data processing engine. Apache Spark is built in the bottom-up mechanism to increase performance. It is faster than Hadoop in terms of speed, mainly for large scale data processing because of in-memory computation and other optimizations.

Apache Flink: It is a platform for distributed computing. It is especially an open-source stream processing engine. It performs well in the case of unordered streaming data. It is easy to use on thousands of nodes with better throughput and latency characteristics.

Apache Storm: It is an open-source platform for real-time distributed computation. Apache Storm is easy to use and set up. It is compatible, scalable and fault-tolerant with any programming language.

H2O: *H2O* is a fast in-memory data processing engine. *H2O* is used for predictive analysis on a large amount of data. It is an open-source, scalable and distributed software that can be implemented on various nodes.

These processing engines are measured on the basis of associated ML tools, latency, supported language, fault tolerance and execution model. Latency is the time duration between initiating a task and receiving the output. Fault tolerance is the mechanism of a system that permits ongoing working appropriately in case of failure of a few modules. Comparative analysis of all the listed data processing engines is represented in table 2 on the basis of several parameters.

Table Error! No text of specified style in document.**:** Comparative analysis of Processing Engines

Data Processing Engine	Current Release (June, 2015)	Supported Languages	Associated ML Tools	Execution Model	In Memory Processing	Fault Tolerance	Low Latency
Haddop	2.7.0	Java	Mahout	Batch	×	✓	×
Spark	1.3.1	Java, R, python, Scala	Mlib, Mahout, H2O	Streaming, Batch	✓	✓	✓
Flink	0.8.1	Java, Scala	SAMOA	Streaming, Batch	✓	✓	✓
Storm	0.9.4	Any	SAMOA	Streaming	✓	✓	✓
H2O	3.0.0.12	Java, Scala, R, Python	H2O, Mahout, Mlib	Batch	✓	✓	✓

Table-2 represents that Apache Spark supports python programming that contains various libraries of machine learning, image processing, etc. Spark performs in-memory processing that provides the result faster. It contains a better ability of fault tolerance and it provides low latency in comparison to other processing engines. Therefore, in the proposed work Apache Spark was used to process the high throughput microarray data.

In the research work, a computer-assisted genomic test method (CAGT) is proposed to predict gene expression levels to reduce breast cancer risk using machine learning and big data analytics.

3 Materials and Methods

In this work carried out, a CAGT (Computer Assisted Genomic Test) is generated on the basis of a clinical report by using ML techniques. Machine Learning is a part of AI, in which a machine learns from its previous experiences. As per Mitchell [27], In ML, machines is used to learn a task on the basis of past experiences and its performance can be evaluated with some parameters like accuracy. Machine learning approaches works in two stages: (i) Analysis of dataset to find out the dependencies of a model. (ii) Output prediction of a model on the basis of projected dependencies. Machine learning is applicable in medical research in numerous uses, where a suitable hypothesis is acquired for a biomedical sample dataset over a multi-dimensional space, using distinctive algorithms [28], [29] [68].

In the work, genes selection is performed using correlation coefficient, and mutual information (ranking method). The explained ratio is used to find the relevant genes for breast cancer. Subsequently, significant genes are selected that has more diagnostic power. The regression technique is applied to predict the gene expressions using the clinical outcome of a patient. Then classification is performed to classify cancer using the predicted gene expression.

The proposed work is implemented on Apache Spark to make the technique fast and scalable using the R programming language and the Sparkler package [30].

3.1 Relevant Gene Selection

Genes are sections of DNA that are passed on the chromosomes and determine specific human characteristics, such as hair color, height and genetic disease. In the research work, commonly established feature selection algorithms are used to locate the most significant genes.

3.1.1 SVM-RFE_MI Gene Selection Technique

SVM-RFE is a wrapper feature selection approach based on SVM. It uses a classification

algorithm to select the features. [31] [69] [70]. The purpose of SVM-RFE is to compute the ranking weights for all features and sort the features according to weight vectors as the classification basis. SVM-RFE removes the least containing features and improves the classification accuracy [32]. SVM- RFE feature selection technique follows three steps, (1) Processing of specific dataset for classification, (2) weight estimation of each feature, (3) the deletion of features with low weight in order to obtain the ranking of features as shown below [32]:

(1) **Input**

- Training Datapoints: X = $[x1, x2, \ldots\ldots, xn]T$
- Label: Y = $[y1, y2, \ldots\ldots, yn]T$
- The complete feature set: S = [1,2,3 $\ldots\ldots n$]
- Reduced feature list: R = []

(2) **Sorting of Selected Features**

- Repetition of procedure until R = [] is received.
- A new training data on the basis of remaining features: X1 = X (: S).
- Classification model used: α = SVM-train (X, y).
- weight Estimation: w =$\sum kxkykxk$
- sorting: $Si = (wi)2$
- Estimate the features containing minimum weight: m = arg min (S).
- Updating the sorted feature list: R = [S(m), R].
- Eliminating the features that contains minimum weight: S = S (1: −1, m + 1: length(S))

3) **Outcome:** This step deals with the sorted feature list. On the basis of prediction accuracy, the feature with the least weight $(wi)2$ is deleted in each iteration. SVM-RFE technique is executed repeatedly until a feature-sorted list is obtained.

In the research work, feature selection is performed on genomic data. Size of the genomic data is 1980 * 24368, where 1980 is sample size and 24368 is number of genes. After applying SVM-RFE, the size of dataset is reduced to 1980 * 120. Then Mutual Information (MI) statistical technique is applied to sort the dataset ranks wise and higher ranked (0.9 to 1) features are extracted. MI sorted list provides 18 breast cancer specific genes among 120 selected genes. These gene selection techniques provide the relevant genes of breast cancer, but most significant genes are required to find that have higher variance and predictable ability.

3.1.2 Explained variance

It is the subpart of total variance which is described by the attributes of the dataset [33]. The maximum percentage of explained variance shows the stronger strength of the attribute. These attributes can make higher predictions of cancer. In the research work, explained variation is calculated by Principal Component Analysis (PCA) because it emphasizes variation in the dataset and provides strong patterns in a dataset. It provides top k genes that have the total explained variance of the dataset. According to the study, acceptable cumulative variance is 70% [34].

In the proposed work, the top 5 genes with 98.5% explained variance is selected as the variance is increased slightly after 5 genes which can result in increased computation.

3.1.3 Principal Component Analysis

PCA incorporates the overall variation in data samples and transforms the original attributes into a smaller set of linear combinations [35]. The smaller set contains the significant details of the dataset. PCA

is commonly used when the goal is to identify the reduced feature set with the highest number of variances, especially when performing multivariate analysis. PCA retains only the first C principal component from total P attributes. PCA is an orthogonal transformation that projects the data from P to C dimensional subspace. In this transformation (P-C) components are vanished which shows that PCA minimizes the variability of data. In the research work, PCA is selected due to the following characteristics [36]:

Principal Component Analysis is intended to maximize the variance of the first C component by minimizing the variance of P-C components. First C components are selected for their highest variance among all the principal components. PCA choose the larger C and these C components have the power to predict the insights of the dataset.

In this proposed work, five components are selected as significant genes. These genes have a 98.5% variance among 18 predictors as shown in table 3. The reduced dataset is trained over SVM classifier and evaluating the performance of the model using SVM prediction accuracy, precision and recall.

3.2 Regression

The regression technique is applied to identify the dependent variable on the basis of multivariate independent variables [37]. In the research work, the gene is a dependent variable whereas clinical outcomes are independent variables. In the proposed work, the least absolute shrinkage and selection operator (lasso) regression technique is used to shrink and remove the coefficients that can reduce variance without increasing the bias. It is especially useful when the dataset has a small number of samples.

3.2.1 Least Absolute Shrinkage and Selection Operator

It is a linear regression approach that has the benefit of shrinkage. [38]. In the shrinkage, samples are shrunk in the direction of the center. LASSO technique delivers the sparse model that contains a reduced feature set. This regression approach is appropriate for multi collinear datasets. It selects the feasible parameters and performs the analysis on these parameters.

Lasso regression technique performs L1 regularization. The loss function is penalized by L1 regularization. The absolute value of the coefficient is incorporated in this penalty, as illustrated in equation [41]:

$$\theta^{lasso} = \min \sum_{i=1}^{n} (y_i - \bar{y})^2 + \lambda \sum_{j=0}^{k} |\theta_j| \quad (1)$$

Here the λ is a regulating factor that determines the severity of the penalty.

LASSO improves model interpretability while increasing prediction accuracy. Therefore, it is the best fit for multivariate regression model [38]. After prediction of gene expression of selected genes, classification technique is applied to classify breast cancer on the basis of these predicted gene expressions.

3.3 Classification

Classification is a machine learning technique that categories the data points in the label according to their type. The medical dataset contains the meaningful biomarkers that help in the categorization of data. In the proposed work, the staking ensemble technique is applied for classification.

3.3.1 Stacking- An Ensemble Technique

Ensemble learning is an ML technique in which numerous models are trained to address the same problem and then come together to get better results. [39]. The main assumption of ensemble learning is that a more accurate and/or robust model obtained when week classifier models are correctly combined.

3.3.2 Week models

Week machine learning models are identified using Bias/Variance trade-off [66]. A low variance and a low bias are the two most essential features of an ML model. The degree of freedom in an ML model should be sufficient to resolve the basic complexities of the data, but not excessive to avoid high variance. It's the well-known tradeoff between bias and variation. Week learner machine learning models exhibit high variance (low degree of freedom) or have a high bias to be reliable (high degree of freedom). Then, ensemble learning is applicable to reduce bias/variance of week learners by combining some of them together to generate a strong learner for better performance. In the proposed research work stacking ensemble, techniques are used.

3.3.3 Stacking

Stacking is one of the ensemble learning techniques [40] that extract the outcome from multiple ML models and combine them to generate a new strong model for better prediction. The ensemble model is applied to assemble the predictions on the test dataset. The key steps of stack ensemble techniques are below given:

- The dataset divides in two parts one is training set and other one is test set.
- The training set further split into 10 subparts:
- A basic classifier model train on the 9 parts and evaluate the performance on the remaining 1 part. This step repeats for each part on training data.
- In this way, Base model is trained on the whole training dataset.
- Now predictions are made using this model on the test set.
- Steps 3 to 5 are then followed by a new base model, which generates a new set of predictions for the train and test sets.
- To create a new model, the predictions from the training set are used as features.
- On the test set, the created base model is used to calculate the final predictions.

4 Experimental Design

4.1 Dataset

In the research work, METABRIC high-throughput sequencing breast cancer dataset [41], [42] is used. Dataset is available in the cBioPortal database. It contains a multi-dimensional dataset of breast cancer. METABRIC dataset has 1,980 data samples that contain both clinical and genomic information among them 548 normal breast tissue samples and 1,432 primary breast tumour samples. Every patient has 27 clinical attributes such as lymph nodes positive, grade, size, age at diagnosis etc. and 24,368 gene expression data. Missing values are imputed using the PC-ImNN imputation technique [43]. Clinical data normalized by min-max normalization [44] in the range of 0 and 1. Table 3 represents the detail of METABRIC breast cancer data.

Table 3: Overall Information of Breast Cancer Dataset

Characteristics	Value	Characteristics	Value
Type of Data Set	Microarray data	**Task Associated**	Classification and Regression
Attribute Type	Real	**Number of Samples**	1,980
Number of genes	24,368	**Clinical Attribute**	27
Cut-off (years)	5	**Diagnosis median age**	61
Short Term Survival (less than 5 years)	491 patients	**Average Survival (Months)**	125.1

4.3 Performance Parameters

In the research work, two performance parameters are applied to assess the performance of regression model and classification model such as: adjusted R-squared and accuracy.

4.3.1 Adjusted R-Squared

In a regression model, it is used to compute the ratio of variance explained by a dependent over the independent variable [48], [70] – [73]. It is a statistical metric that considers the significant predictors that precisely affect the dependent variables. As a result, in multivariate regression models, it performs well. It is computed as follows [48]:

$$\bar{R}^2 = 1 - (1 - R^2)\left[\frac{s-1}{s-(v+1)}\right] \quad (2)$$

where, V signifies the total independent variable and total number of samples is denoted by the letter S. The adjusted R-Squared value has the range between 0 to 1 like $0 \leq R\bar{\wedge}2 \leq 1$. If the value of $\bar{R}^2$ is close to 1, it indicates that the predicted regression line is the same as the actual regression line [74] – [78].

4.3.2 Accuracy

According to the International Organization of Standardization (ISO), accuracy is the trueness of the model [49]. It is a combination of both kinds of systematic observational and random error, so high accuracy requires both high trueness and high precision [79] [80]. It was used to calculate the proportion of samples that were correctly categorized [79].

$$accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad (3)$$

In eq. (3), TP, TN, FP, and FN are True Positive, True Negative, False Positive and False Negative respectively [81] - [83]. Accuracy is calculated as sum of true positives, true negatives divided by true positives, true negatives, false negatives, and false positives. For good classifiers, true negative rate and true positive rate both should be nearer to 100%.

5 Results and Discussion

The findings of the proposed experiments are described in this section. A gene selection approach SVM-RFE-MI is designed to find the significant genes associated with breast cancer. The designed technique for breast cancer is improved by combining SVM-RFE, Explained Variance and Mutual Information rank method. SVM-RFE is used to eliminate non-functional genes and the ML method is used to get the rank of gene between 0-1. Explained Variance using PCA is applied to find the most significant genes.

In the experiment, SVM-RFE-MI provides 18 significant genes from a dataset containing 24368 genes. The model is again improved to identify genes containing higher explained variance using PCA which provides the top 5 significant genes as shown in Table 4. These 5 selected genes are containing 98.5% of variance which is in an acceptable range.

Table 4: Top 5 genes have 98.5% explained variance ratio

Gene	Explained Variance	Cumulative Variance	Description
PIK3CA	38	38	Regulation of hormones and maturation of cells
TP53	34	72	Guardian of the genome
GATA3	14	86	Independent prognostic marker
AURKA	6.5	92.5	Serine-threonine kinases
PTEN	6	98.5	Tumor suppressor gene

The brief description of each gene is shown below.

- PIK3CA: It is the recurrently mutated gene in breast cancer [50]. PIK3CA provides the instruction to generate p110 alpha protein and add a cluster of oxygen to perform the action of the P13KCA gene. The actions of P13K are to send signals for many cell activities, including migration (movement) of cells, cell proliferation (division) and growth, the passage of the materials within cells, the creation of new proteins, and cell survival.

- TP53: It is also known as the guardian of the genome. The TP53 gene gives instructions for making the protein p53. [51]. p53 protein works for tumour suppressors. TP53 regulates cell division by preventing cells from proliferating (dividing) and speedily developing in an uncontrolled way. All cells in the body contain the p53 protein, which binds to DNA directly in the nucleus [52].

- GATA3: It is used to regulate the expression of a variety of biologically and clinically significant genes. [53]. As per the recent study, GATA-3 has been linked to positive breast cancer pathologic features such as positive estrogen receptor (ER) and negative lymph node status, according to a recent study. It's also been discovered to be a standalone prognostic marker, with a low expression indicating a higher probability of recurrence of breast cancer [54].
- AURKA: It is a kinase, which is significant for cell division. AUREKA's major role is to control chromosomal segregation during mitosis. AURKA kinase mutations cause cell division failure and cellular progression to be harmed [55] [83].

- PTEN: This gene encodes the protein called Phosphatase and tensin homolog (PTEN). Mutation in the PTEN gene leads to many cancers along with breast cancer [56]. It works as a tumour suppressor gene with the help of phosphatase protein. This protein prevents

cells from dividing and growing rapidly and regulates the cell cycle. It is a target for various anticancer drugs [57] [84] [85].

These selected genes are verified from Genetics Home Reference and Cancer Genetics Web [58], [59]. These genes are directed for breast cancer detection by researchers. Identified genes have higher productiveness and variance for breast cancer prediction at an early stage. Selected genes contain less correlation with each other. The relation between each gene is represented in Table 5. As a result, if experts target these genes, cancer will be detected at an earlier stage.

Table 5: The top 5 genes' correlation matrix

Gene	PIK3CA	TP53	GATA3	AURKA	PTEN
PIK3CA	1	0.26	0.31	-0.12	0.41
TP53		1	0.19	0.28	0.21
GATA3			1	0.37	0.31
AURKA				1	-0.03
PTEN					1

According to Table 5, these genes are less correlated. So, the prediction is performed individually on each gene. In the work carried out, the multivariate LASSO regression technique is used to predict the expression of each gene. In the Lasso model, each gene is considered as the dependent variable and all the attributes of the clinical dataset are considered as an independent variable. The results of LASSO model on the basis of $\bar{R}2$ are represented in table 6. Results show the performance of the Lasso regression model in terms of $\bar{R}^2$. The LASSO regression technique has a good prediction accuracy as it proceeds with independent attributes.

Table 6: Estimation of predicted genes using LASSO regression technique on the basis of $\bar{R}2$

Gene	$\bar{R}^2$
PIK3CA	0.95
TP53	0.98
GATA3	0.87
AURKA	0.89
PTEN	0.92

As per the Table 6, LASSO gives minimum 0.13 error for GATA3 gene, according to the Study [64], $\bar{R}^2$ should be near to one and greater than 0.7. It indicates that the model's $\bar{R}^2$ error is acceptable. Now the classification is performed on the basis of these predicted gene expressions.

In the research work, a stacking ensemble technique is used to classify the breast cancer stage on the basis of predicted gene expression. In this stacking technique, random forest [60], K-Nearest Neighbor (KNN) [61], Support Vector Classifier (SVC) [62] and multilayer perceptron (MLP) [63] machine learning techniques are used to ensemble [67]. Table 7 shows the classification accuracy, precision and recall of ensemble techniques and other machine learning techniques.

Table 7: A comparative analysis of classification techniques to classify predicted genes

	Classification Technique				
	Random Forest	**KNeighbors**	**SVC**	**MLP**	**Ensemble**
Accuracy	0.90	0.89	0.91	0.95	0.98
Precision	0.81	0.80	0.88	0.88	0.89
Recall	0.48	0.47	0.49	0.51	0.52

Results show that predicted gene expressions are able to diagnose breast cancer to some extent. It will be beneficial for those patients who have chances to get breast cancer in future or run with the financial crises. Figure 1 represents the graphical representation of classification results.

Fig 1: Performance estimation of cancer classification on the basis of accuracy

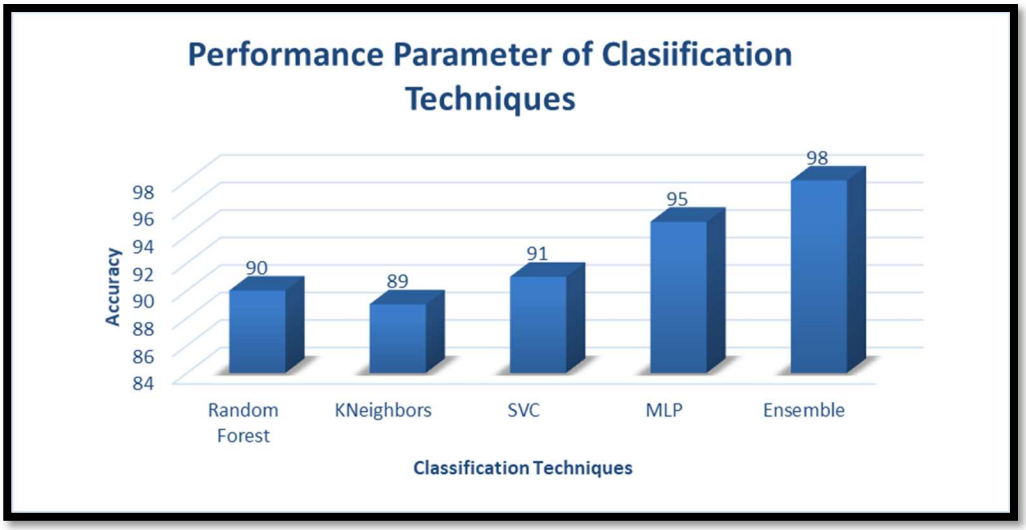

Figure 1 shows that the ensemble technique outperformed the other individual machine learning technique in terms to classify predicted gene expressions.

6 Conclusion

The most common malignancy among women is breast cancer. It occurs when breast cells start to grow abnormally because of the gene mutation. A genomic test is preferred to determine the gene mutation at the primary stage, but the test is time-consuming and expensive in underdeveloped nations. In the work carried out, a novel computer-assisted genetic test method is proposed to predict the gene expression in reduced time and cost. The proposed method provides the expression of significant genes and diagnoses the cancer stage. The model helps to categorize the samples in benign and malignant cancer along this reduces the risk of breast cancer by identifying the gene mutations at the primary stage. In the computer-assisted genetic test method, the SVM-RFE-MI approach is proposed for gene selection. LASSO regression technique is used to predict gene expression, followed by the stacking ensemble strategy to categorize cancer. Prediction of gene expression is evaluated using adjusted R-Squared and classification of cancer by accuracy performance parameters. As per the results, adjusted R-Squared is found to be within the standard acceptable range and ensemble techniques outperform other machine learning approaches in terms of accuracy. It signifies that the test method provides better gene prediction. The proposed technique will provide reports proximately after clinical outcome with no cost. It is helpful for patients who are suffering from breast cancer. The technique will help to reduce the mortality rate by diagnosing cancer at an early stage.

The work can be beneficial for the performance and stability analysis of ensemble feature selection for cancer prediction. Trials can explore other applications and related datasets in the prediction domain. Heterogeneous ensemble gene selection technique and other similarity measures can be used in future experiments. Advance ensembles can also be discovered using hybrid techniques. Protein data can be included to get the disease prediction at an increased depth.

Acknowledgements

Not applicable.

Author's Contributions

The work is done by equal cooperation of authors.

Funding

This research is supported by Symbiosis International University, Pune, Maharashtra, India

Competing Interests

The authors declare that they have no competing interests.

References

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[2] Breast Cancer Statistics, "Globocan Project." http://www.breastcancerindia.net/statistics/stat_global.html (accessed Mar. 05, 2019).

[3] R. Sarin, "The cost of genetic testing for cancer has to come down," 2018. https://www.livemint.com/Politics/LSN7wtUjRj3iR0ZDk5ncZO/The-cost-of-genetic-testing-for-cancer-has-to-come-down.html (accessed Dec. 20, 2018).

[4] J. D. Hoheisel, "Microarray technology: Beyond transcript profiling and genotype analysis," *Nature Reviews Genetics*, vol. 7, no. 3, pp. 200–210, 2006, doi: 10.1038/nrg1809.

[5] T. Zeng and J. Liu, "Mixture classification model based on clinical markers for breast cancer prognosis," *Artificial Intelligence in Medicine,* vol. 48, no. 2–3, pp. 129–137, 2010, doi: 10.1016/j.artmed.2009.07.008.

[6] L. T. Scaria and T. Christopher, "A Bio-inspired Algorithm based Multi-class Classification Scheme for Microarray Gene Data," *Journal of Medical Systems,* vol. 43, no. 7, 2019, doi: 10.1007/s10916-019-1353-y.

[7] M. Jansi Rani and D. Devaraj, "Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification," *Journal of Medical Systems*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1372-8.

[8] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002, doi: 10.1038/ng765.

[9] P. Ferreira, I. Dutra, R. Salvini, and E. Burnside, "Interpretable models to predict Breast Cancer," Proceedings - 2016 *IEEE International Conference on Bioinformatics and Biomedicine, BIBM* 2016, pp. 1507–1511, 2017.

[10] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent and M. Bittner, "Multivariate measurement of gene expression relationships," *Genomics*, vol. 67, no. 2, pp. 201–209, 2000, doi: 10.1006/geno.2000.6241.

[11] S. Muro, I. Takemasa, S. Oba, R. Matoba, N. Ueno, C. Maruyama, R. Yamashita, M. Sekimoto, H. Yamamoto, S. Nakamori and M. Monden, "Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data," *Genome biology*, vol. 4, no. 3, 2003, doi: 10.1186/gb-2003-4-3-r21.

[12] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A gene signature for breast cancer prognosis using support vector machine," *2012 5th International Conference on Biomedical Engineering and Informatics, BMEI* 2012, no. October, pp. 928–931, 2012, doi: 10.1109/BMEI.2012.6513032.

[13] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 928–937, 2015, doi: 10.1109/TCBB.2014.2377729.

[14] C. Kandaswamy, L. M. Silva, L. A. Alexandre, and J. M. Santos, "High-Content Analysis of Breast Cancer Using Single-Cell Deep Transfer Learning," *Journal of Biomolecular Screening*, vol. 21, no. 3, pp. 252–259, 2016.

[15] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. Kahn, and E. S. Burnside, "Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration," *Cancer*, vol. 116, no. 14, pp. 3310–3321, 2010.

[16] Z. Zhou and Y. Jiang, “Medical Diagnosis With C4 . 5 Rule Preceded by,” vol. 7, no. 1, pp. 37–42, 2003.
[17] Z.-H. Zhou, “Data mining techniques: To predict and resolve breast cancer survivability,” *Scholarpedia*, vol. 4, p. 2776, 1990.
[18] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng and K. Li, “A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 919–933, 2017, doi: 10.1109/TPDS.2016.2603511.
[19] T. Shaikhina and N. A. Khovanova, “Handling limited datasets with neural networks in medical applications: A small-data approach,” *Artificial Intelligence in Medicine*, vol. 75, pp. 51–63, 2017, doi: 10.1016/j.artmed.2016.12.003.
[20] J. Listgarten, J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour and A. Driga, “Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms,” *Clinical Cancer Research*, vol. 10, no. 8, pp. 2725–2737, 2004.
[21] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, “Robust predictive model for evaluating breast cancer survivability,” *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, pp. 2194–2205, 2013.
[22] C. Park, J. Ahn, H. Kim, and S. Park, “Integrative gene network construction to analyze cancer recurrence using semi-supervised learning,” *PLoS ONE*, vol. 9, no. 1, pp. 1–9, 2014, doi: 10.1371/journal.pone.0086309.
[23] J. Kim and H. Shin, “Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data,” *Journal of the American Medical Informatics Association*, vol. 20, no. 4, pp. 613–618, 2013.
[24] W. Kim, K.S. Kim, J.E. Lee, D.Y. Noh, S.W. Kim, Y.S. Jung, M.Y. Park and R.W. Park, “Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine,” vol. 15, no. 2, pp. 230–238, 2012, doi: 10.4048/jbc.2012.15.2.230.
[25] O. Gevaert, F. de Smet, D. Timmerman, Y. Moreau, and B. de Moor, “Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks,” *Bioinformatics*, vol. 22, no. 14, pp. 184–190, 2006.
[26] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, “LIBLINEAR: A Library for Large Linear Classification Rong-En Fan Xiang-Rui Wang,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, [Online]. Available: http://www.csie.ntu.edu.tw/.
[27] T. M. Mitchell, *The Discipline of Machine Learning*, vol. 17, no. July. 2006.
[28] F. Cabitza, R. Rasoini, and G. F. Gensini, “Unintended consequences of machine learning in medicine,” *JAMA - Journal of the American Medical Association*, vol. 318, no. 6, pp. 517–518, 2017, doi: 10.1001/jama.2017.7797.
[29] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/NEJMra1814259.
[30] “Sparklyr,” 2016. [Online]. Available: https://blog.rstudio.com/2016/09/27/sparklyr-r-interface-for-apache-spark/.
[31] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, “SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–18, 2018.
[32] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, “Gene selection for tumor classification using neighborhood rough sets and entropy measures,” *Journal of Biomedical Informatics*, vol. 67, pp. 59–68, 2017, doi: 10.1016/j.jbi.2017.02.007.
[33] J. A. Rosenthal, Statistics and data interpretation for social work. 2012.
[34] “Acceptable cumulative explained variance,” 2018. [Online]. Available: https://support.sas.com/publishing/pubcat/chaps/55129.pdf.
[35] H. Todorov, D. Fournier, and S. Gerber, “Principal components analysis: theory and application to gene expression data analysis,” *Genomics and Computational Biology*, vol. 4, no. 2, p. 100041, 2018.
[36] R. Cheplyaka, “PCA-Explained Variance,” 2017. https://ro-che.info/articles/2017-12-11-pca-explained-variance (accessed Mar. 28, 2019).
[37] S. Chatterjee and A. S. Hadi, Regression analysis by example, 5th ed. Canada: John Wiley & Sons, 2015.
[38] A. S. Dalalyan, M. Hebiri, and J. Lederer, “On the prediction performance of the Lasso,” *Bernoulli*, vol. 23, no. 1, pp. 552–581, 2017, doi: 10.3150/15-BEJ756.
[39] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” Frontiers of Computer Science, vol. 14, no. 2, pp. 241–258, 2020, doi: 10.1007/s11704-019-8208-z.
[40] A. Ekbal and S. Saha, “Stacked ensemble coupled with feature selection for biomedical entity extraction,” *Knowledge-Based Systems*, vol. 46, pp. 22–32, 2013, doi: 10.1016/j.knosys.2013.02.008.
[41] “METABRIC Breast Cancer Dataset,” Nat Commun, 2016. https://www.cbioportal.org/study/summary?id=brca_metabric.
[42] I. Rezaeian, E.J. Mucaki, K. Baranova, H.Q. Pham, I. Rezaeian, D. Angelov, A. Ngom, L.Rueda, and P.K. Rogan, “Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning [version 1; referees: 2 approved with reservations],” F1000Research, vol. 5, no. May, pp. 1–24, 2016, doi: 10.12688/F1000RESEARCH.9417.1.
[43] M. Gupta and B. Gupta, “A new scalable approach for missing value imputation in high-throughput microarray data on apache spark,” *International Journal of Data Mining and Bioinformatics*, vol. 23, no. 1, pp. 79–100, 2020, doi: 10.1504/IJDMB.2020.105438.
[44] S. G. K. Patro and K. K. sahu, “Normalization: A Preprocessing Stage,” *Iarjset*, pp. 20–22, 2015.

[45] “Invasive Ductal carsinoma (IDC),” 2019. https://www.breastcancer.org/symptoms/types/idc (accessed Apr. 04, 2020).
[46] K. G. Srinivasa, S. G. M., and S. H., “Apache Spark,” pp. 73–83, 2018, doi: 10.1007/978-3-319-77800-6_4.
[47] X. Meng, X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D.B. Tsai, M. Amde, S. Owen and D. Xin, “MLlib : Machine Learning in Apache Spark,” vol. 17, pp. 1–7, 2016, doi: 10.1145/2882903.2912565.
[48] V. Rousson and N. F. Goşoniu, “An R-square coefficient based on final prediction error,” *Statistical Methodology*, vol. 4, no. 3, pp. 331–340, 2007, doi: 10.1016/j.stamet.2006.11.004.
[49] S. García, A. Fernández, J. Luengo, and F. Herrera, “A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability,” *Soft Computing*, vol. 13, no. 10, pp. 959–977, 2009.
[50] G. Conditions, “PIK3CA Genetics Home Reference,” 2020. https://ghr.nlm.nih.gov/gene/PIK3CA (accessed Apr. 05, 2020).
[51] J. Y. Wei Liu and Margo MacDonald, “Mutational processes shape the landscape of TP53 mutations in human cancer,” *Physiology & behavior*, vol. 176, no. 1, pp. 139–148, 2016, doi: 10.1016/j.physbeh.2017.03.040.
[52] National Institutes of Health, “Genetics Home Reference TP53 gene,” 2017. [Online]. Available: https://ghr.nlm.nih.gov/gene/TP53.
[53] D. Voduc, M. Cheang, and T. Nielsen, “GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value*,” Cancer Epidemiology Biomarkers and Prevention*, vol. 17, no. 2, pp. 365–373, 2008, doi: 10.1158/1055-9965.EPI-06-1090.
[54] N. Emmanufl, K. A. Lofgren, E. A. Petersion, D. R. Meier, E. H. Jung, and P. A. Kenny, “Mutant GATA3 Actively Promotes the Growth of Normal and Malignant Mammary Cells,” *Physiology & behavior*, vol. 176, no. 1, pp. 139–148, 2016, doi: 10.1016/j.physbeh.2017.03.040.
[55] H. J. Donnella, J.T. Webber, R.S. Levin, R. Camarda, O. Momcilovic, N. Bayani, K.N. Shah, J.E. Korkola, K.M. Shokat, A. Goga and J.D. Gordan, “Kinome rewiring reveals AURKA limits PI3K-pathway inhibitor efficacy in breast cancer,” *Nature Chemical Biology*, vol. 14, no. 8, pp. 768–777, 2018, doi: 10.1038/s41589-018-0081-9.
[56] Y. R. Lee, M. Chen, and P. P. Pandolfi, “The functions and regulation of the PTEN tumour suppressor: new modes and prospects,” *Nature Reviews Molecular Cell Biology*, vol. 19, no. 9, pp. 547–562, 2018,
[57] National Institutes of Health, “Genetics Home Reference PTEN gene,” 2015. [Online]. Available: https://ghr.nlm.nih.gov/gene/PTEN.
[58] “US National Library of Medicine,” Genetic Home Reference, 2020. https://ghr.nlm.nih.gov/gene.
[59] “Cancer Genetics Web,” 2017. http://www.cancerindex.org/geneweb/X0401.htm (accessed Nov. 10, 2018).
[60] R. Díaz-Uriarte and S. Alvarez de Andrés, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, pp. 1–13, 2006, doi: 10.1186/1471-2105-7-3.
[61] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, “An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm,” pp. 280–285, 2015, doi: 10.12792/iciae2015.051.
[62] M. Cinelli, Y. Sun, K. Best, J.M. Heather, S. Reich-Zeliger, E. Shifrut, N. Friedman, J. Shawe-Taylor and B. Chain, “Feature selection using a one-dimensional naïve Bayes’ classifier increases the accuracy of support vector machine classification of CDR3 repertoires,” *Bioinformatics*, vol. 33, no. 7, pp. 951–955, 2017, doi: 10.1093/bioinformatics/btw771.
[63] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, “Multilayer Perceptron: Architecture Optimization and Training,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, p. 26, 2016.
[64] K. Y. Kim, J. Park, and R. Sohmshetty, “Prediction measurement with mean acceptable error for proper inconsistency in noisy weldability prediction data,” *Robotics and Computer-Integrated Manufacturing*, vol. 43, pp. 18–29, 2017.
[65] M. Gupta and B. Gupta., “Survey of Breast Cancer Detection Using Machine Learning Techniques in Big Data”, *Journal of Cases on Information Technology (JCIT),* vol. 21, no. 3, pp.80-92, 2019.
[66] Weak Learner, “Ensemble methods: bagging, boosting and stacking”, https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stackingc9214a10a205 (accessed on Apr. 2020).
[67] Xiong, Hu, Chuanjie Jin, Mamoun Alazab, Kuo-Hui Yeh, Hanxiao Wang, Thippa Reddy Reddy Gadekallu, Weizheng Wang, and Chunhua Su. "On the design of blockchain-based ECDSA with fault-tolerant batch verication protocol for blockchain-enabled IoMT." *IEEE Journal of Biomedical and Health Informatics*, 2021.
[68] C. Iwendi, S. Khan, J. H. Anajemba, A. K. Bashir and F. Noor, “Realizing an Efficient IoMT-Assisted Patient Diet Recommendation System Through Machine Learning Model,” in *IEEE Access*, vol. 8, pp. 28462-28474, Jan. 2020
[69] A. Makkar, M. S. Obaidat and N. Kumar, "FS2RNN: Feature Selection Scheme for Web Spam Detection Using Recurrent Neural Networks," in *IEEE Global Communications Conference (GLOBECOM),* pp. 1-6, Dec. 2018.
[70] P. Kumar, R. Kumar, G. Srivastava, G.P. Gupta, R. Tripathi, T.R. Gadekallu and N. N. Xiong, "PPSF: A Privacy-Preserving and Secure Framework Using Blockchain-Based Machine-Learning for IoT-Driven Smart Cities," in *IEEE Transactions on Network Science and Engineering,* vol. 8, no. 3, pp. 2326-2341, Sept. 2021.
[71] A. Makkar, N Kumar, AY. Zomaya, S. Dhiman, “SPAMI: A cognitive spam protector for advertisement malicious images,” *Information Sciences*. Vol. 1, no. 54, pp.17-37, Nov. 2020.
[72] A. Makkar, U. Ghosh, D. B. Rawat and J. Abawajy, "FedLearnSP: Preserving Privacy and Security using Federated Learning and Edge Computing," in *IEEE Consumer Electronics Magazine*, doi: 10.1109/MCE.2020.3048926.

[73] A. Makkar, U. Ghosh, P. K. Sharma and A. Javed, "A Fuzzy-based approach to Enhance Cyber Defence Security for Next-generation IoT," in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2021.3053326.

[74] A. Makkar, U. Ghosh and P. K. Sharma, "Artificial Intelligence and Edge Computing-Enabled Web Spam Detection for Next Generation IoT Applications," in *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25352-25361, Nov. 2021.

[75] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar and J. J. P. C. Rodrigues, "Tactile Internet for Smart Communities in 5G: An Insight for NOMA-Based Solutions," in *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3104-3112, May 2019.

[76] I. Budhiraja, "A Systematic Review on NOMA Variants for 5G and Beyond," in *IEEE Access*, vol. 9, pp. 85573-85644, 2021,

[77] I. Budhiraja, N. Kumar and S. Tyagi, "ISHU: Interference Reduction Scheme for D2D Mobile Groups Using Uplink NOMA," in *IEEE Transactions on Mobile Computing*, doi: 10.1109/TMC.2021.3051670

[78] I. Budhiraja, N. Kumar and S. Tyagi, "Energy-Delay Tradeoff Scheme for NOMA-Based D2D Groups With WPCNs," in *IEEE Systems Journal,* vol. 15, no. 4, pp. 4768-4779, Dec. 2021.

[79] I. Budhiraja, N. Kumar and S. Tyagi, "Deep-Reinforcement-Learning-Based Proportional Fair Scheduling Control Scheme for Underlay D2D Communication," in *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3143-3156, Mar. 2021

[80] M. Babar, M. S. Khan, U. Habib, B. Shah, F. Ali and D. Song, "Scalable Edge Computing for IoT and Multimedia Applications Using Machine Learning", Human-*centric Computing and Information Sciences*, *11, 2021.*

[81] M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index", *Human-centric Computing and Information Sciences*, vol. *7*, no. 4, pp.1-13, 2017.

[82] M. Elhoseny, G. Ramírez-González, O.M. Abu-Elnasr, S.A. Shawkat, N. Arunkumar and A. Farouk, "Secure medical data transmission model for IoT-based healthcare systems" *IEEE Access*, vol. 6, pp.20596-20608, 2018.

[83] N. Rifi, N. Agoulmine, N. Chendeb Taher and E. Rachkidi, "Blockchain technology: is it a good candidate for securing iot sensitive medical data?", *Wireless Communications and Mobile Computing*, *2018*.

[84] W. Wang, H. Xu, M. Alazab, T. R. Gadekallu, Z. Han and C. Su, "Blockchain-Based Reliable and Efficient Certificateless Signature for IIoT Devices," in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3084753.

[85] Balamurugan, N. M., et al. "DOA tracking for seamless connectivity in beamformed IoT-based drones." Computer Standards & Interfaces 79 (2022): 103564.

[86] W. Wang *et al*., "Blockchain and PUF-based Lightweight Authentication Protocol for Wireless Medical Sensor Networks," in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2021.3117762.