OPEN ACCESS

GAVIN PUBLISHERS

Advances in Proteomics and Bioinformatics

Research Article

Patel N, et al. Adv Proteomics Bioinform 03: 115.

DOI: 10.29011/2690-0092.100015

Improved Prediction of Breast Cancer on Epigenomics Data using Feature Selection and Machine Learning

Nilisha Patel[1], Kalpdrum Passi[1*], Chakresh Kumar Jain[2]

[1]Department of Mathematics and Computer Science, Laurentian University, Sudbury, Canada

[2]Department of Biotechnology, Jaypee Institute of Information Technology, India

[*]**Corresponding author:** Kalpdrum Passi, Department of Mathematics and Computer Science, Laurentian University, Sudbury, Canada

Citation: Patel N, Passi K, Jain CK. (2020) Improved Prediction of Breast Cancer on Epigenomics Data using Feature Selection and Machine Learning, Adv Proteomics Bioinform 03: 115. DOI: 10.29011/2690-0092.100015

Received Date: 07 March, 2020; **Accepted Date:** 27 March, 2020; **Published Date:** 31 March, 2020

Abstract

Cancer is a complex disease associated with the alteration in DNA methylation, histone modification, post transcriptional modifications with epigenetic changes in genes and pathways at different level, leading to generate the malignant phenotypes. However, with plethora of datasets for epigenetic status an accurate statistical insight is obtained which could provide the shreds of evidences for phenotypic alteration and certain behavior of DNA is still missing in breast cancer. In this paper, four different types of data viz. Methylation, Histone, Human Genome and RNA-Seq data have been used to discriminate between cancerous and non-cancerous cells in breast cancer. The data has been pre-processed with in-house developed R-Script and The Weka tool deployed for the feature selection and classification. Four different types of feature selection methods viz., PCA, Gain Ratio, ReliefF, and CFS and eight different types of classifiers with 10-fold cross validation have been used for accurate classification of data. Different combinations/ratio of data sizes for training and testing were used for achieving optimized classification accuracy.

The entire data processing is done for the prediction and testing of breast cancer. With the help of the machine learning method, the epigenetics data shows the prediction of breast cancer in the given set of cells. The results are supported by various data tables and graphs which demonstrate how CFS is the best feature selection method and linear SVM is the best classifier out of all the selected ones since it gives 100% accuracy the greatest number of times with the combination of different feature selection methods. However, for the CFS feature selection, Random Forest classifier selects the best model for different characteristics giving 100% accuracy the greatest number of times. The total number of features is 1564 spanning the four categories of CpG methylation, Histone H3 data, nucleotide composition and RNA-seq data. CFS feature selection method selects the best 245 features. CFS and Linear SVM classification technique selects the best overall model with various characteristics, with accuracy of 100% using 10-fold cross-validation. With the help of various tests and methods, it was observed that different classifiers exhibit different results when tested under various training-testing data. Also, the outcome varies with the variations in dataset and data processing techniques. However, feature selection method plays a vital role in driving the results. With the help of various tests and methods, it was observed that different classifiers exhibit different results

Index Terms: Epigenomics; Histone; DNA Methylation; Human Genome; RNA-Sequencing; Feature Selection

Introduction

Plenty of diseases are diagnosed that observe epigenetic alterations and thus, modify the expressions of genes. Cancer is one such disease known to witness some of the critical epigenetic phenomena. Gradually spreading its wings, cancer has so far victimized quite a huge mass of people. Lacking enough background information like causes, symptoms, and cure, cancer is today one of the deadliest diseases [19]. For the early diagnosis of breast cancer detection, several researches are on-going at many laboratories and research centres across the globe. Researchers are striving hard to collect sufficient quantitative methods that support the contribution of epigenetic data in identifying the onset as well as the spread of cancer in patients. Datasets of cancerous cells are available on

various sites so that researchers can process the data and conclude certain epigenetic studies.

Medical science and machine learning together have stretched the horizons of data processing of epigenetic details for cancer. The emerging data from high-throughput technologies for DNA Methylation, Human Genome and Histone are useful in acquiring a better understanding of epigenetic alterations and expression of genes. With the combination of such datasets and data processing techniques, tools can be designed that can provide accurate bioinformatics results. Out of all other epigenetic regulations, histone modification and DNA methylation are the most crucial ones. In the region of protein-coding genes, CpG methylation is observed. However, in certain intergenic regions also, such alterations are noticed. It has been concluded that CpG methylation occurs in promoters located upstream of the transcription starting site, and increased methylation in the promoter is negatively associated with the gene association level [11].

The datasets have records of increased methylation (hyper methylation) as well as reduced methylation (hypo methylation). Such types of DNA methylation are known to alter the activities of the DNA segment without modifying its sequence. Typically, methylation represses the gene transcription thereby affecting the key DNA activities [9, 17]. The most common results of such activities are X-chromosome inactivation, genomic imprinting, carcinogenesis, and aging. Out of the four bases of DNA, adenine and cytosine can be methylated. For cancer affected cells, the gene promoter CpG Island goes through hyper methylation. This alteration is one of the most critically observed activities in cancer cells [1, 11]. As cancer progresses, genes are gradually activated or silenced. Such silencing occurs at multiple CpG locations in the CpG Island. Budding methods like Illumina Infinium provide a better analysis of CpG methylation [18]. Another crucial epigenetic modification is Histone modification. Histones are the core of nucleosomes. DNA sequences wrap around these nucleosomes [23]. A certain portion of methylation, as well as acetylation, can be observed in histones, nevertheless not all such activities are subjected to cancer. Such activities can either open up the local chromatin structure to enable the gene expression or else close the structure to block the gene expression. Out of all Histones, Histone 3 serves as the most suitable one for measuring the gene expression status [12].

A large amount of histone alteration data can be acquired from CHIP-Sequencing. Epigenetic data is, however, insufficient to mark whether such alteration will result in up-regulation or down-regulation of gene expression. For predicting the complete analysis, classifiers are used that support the machine learning technique. A huge set of histone modification data, methylation data, human genome data, and RNA-Seq are examined using this model [4]. The processing is done on the publically available data sets of breast cancer. Tabl et al. [21] proposed an approach to identify relevant genes for breast cancer survivability on specific therapies. Supervised and unsupervised machine learning approaches were used to solve a multi-classification issue in which samples are classified on the basis of a combination of 5-year survival and treatment focussing on hormone therapy, radiotherapy and surgery. In a subsequent paper, Tabl et al. [22] proposed a machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. They presented a hierarchical machine learning system that predicts the 5-year survivability of the patients who underwent specific therapy. They define classes for a combination of survivability and the given therapy.

Software and tools

Data processing is performed using in-house developed R-Script whereas; classification, as well as feature selection, is done using data mining software - Weka.

Methods

In the context of breast cancer, an approach is introduced here that can manifest the cancerous cells and discriminate them from non-cancerous cells. Several data sets are used for feature selection and classification purposes. The data sets include methylation, human genome, histone, and RNA-Sequencing. These data of breast cancer are processed further using smart machine learning tools and algorithms.

DNA Methylation Data

Taken from Illumina's Infinium Human Methylation 450 Bead chip, The Cancer Genome Atlas Methylation data has been collected and further used to extract the features of CpG methylation for breast cancer [3]. The exons, coordinates of CpG and coding sections are selected from the available Illumina file. Data is processed such that comprehensive information can be collected from the sets.

Histone Data

H3k4me3, H3k27me3, and H3k36me3 are the three Histone marker CHIP-Seq data sets that are used for the analysis. These data sets are potential enough to provide the actionable targets in case of the treatment of breast cancer [6, 13]. These are evident resources of the modification profile of histone to identify the epigenetic landscapes for the cells of breast cancer. To normalize the collected data over the total quantity of reads, R-Script is used.

Human Genome Data

Hg19 genome FASTA files were downloaded to extract the data of nucleotide composition. The PhastCons46Way scores were taken from the UCSC genome browser [15]. The conservation scores are considered for species like placental animals, vertebrates and primates. Further R-Script was used to extract features based on these scores.

RNA-Seq Data

RNA sequencing reveals the quantity of RNA in a given sample. Digital data, in terms of aligned read-counts are obtained which in-turn provides a dynamically wide range of results that can help improve the detection sensitivity for the transcripts.

Additionally, it is a cost-effective method that provides in-depth knowledge of the RNA profile. While calculation for the gene level expression of various transcripts map to a single Ref-Seq ID, the results are calculated based on the geometric mean of differential expressions. However, if the read count is zero, the counts are increased by one for all the transcripts. Later, the expression is expressed as binary data, either down-regulated or up-regulated. The data for RNA-Seq gene expression has been collected from the TCGA Research Network: http://cancergenome.nih.gov

Feature Extraction

Using these diverse data sets, various features are obtained. All these features are then extracted using exclusive methods. Histone modification data, DNA sequence data, and DNA methylation 450K data are combined as predictors. Later, RNA-Seq expression data (up versus down-regulated genes put into binary) is used as response variables. Illumina 450K annotation file is used to get the specific features for breast cancer cells.

CpG Methylation features

Differential expressions of the methylated CpG sites were processed using the limma library in R. Specifically, the function toptable was used to determine the log fold change (logFC) between the cancer and normal tissues as well as the average methylation (avgMval) of each CpG site across the two types of tissues [20]. A positive logFC indicates hypermethylation whereas a negative logFC indicates hypomethylation [2]. Additional segment-based features were also considered. These include the number of hypermethylated (numHyper) and hypomethylated probes (numHypo) on a segment of a given transcript. For example, first_exon_numHyper refers to the number of hypermethylated probes on the first exon. Two other types of features are the average of logFC and avgMval of all CpG probes on a segment of the transcript, e.g. the average logFC of all probes on the first exon of a given transcript (first_exon_avglogFC).

Special effort was paid to compute distances of CpG probes to exon-exon junctions. Given that one or more CpG sites may exist on the individual exon segments of a transcript (including the first and last exons), transcript-level maximum, minimum and average distances of any hyper/hypo-methylated probe to the nearest 5' or 3' exon-exon junction were computed (maxHypoTo5, minHypoTo5, avgHypoTo5, maxHypoTo3, minHypoTo3, avgHypoTo3, maxHyperTo5, minHyperTo5, avgHyperTo5, maxHyperTo3, minHyperTo3, and avgHyperTo3) [10].

Histone marker change feature

After the alignment of raw histone marker data, the aligned histone marker reads were intersected with the segments of each transcript using the multicov function from the BEDTools package [16]. The histone reads were then normalized per 1000 bp length of each segment per 1 million aligned read library. Similar to the CpG methylation features, the histone marker modification features were extracted on a segment-by-segment basis. Initials are used to represent the individual cell lines where the features come from: A for the MCF-7 cell line and S for the SAEC cell line. Following the initial is a number representing the specific histone H3 methylation marker: 4 for H3k4me3, 27 for H3k27me3, and 36 for H3k36me3. As a result, features are named as segment_ cell type and histone modification type (e.g. first_exon_A4). In order to compare histone modification between the cancer and non-cancer cell types, the differences of the reads between them were divided by the average of the two (e.g. a feature named first_ exon_A4_minus_S4_divavg) [10].

Nucleotide feature

In each segment of the transcript, four different types of nucleotide features were extracted: single nucleotide composition, dinucleotide composition, trinucleotide composition, and the length of each segment. Nucleotide sequences of Hg19 reference genome were processed using the Biostrings library in R [14].

Conservation feature

Conservation score per segment was calculated as the arithmetic mean of the conservation score per nucleotide in that segment. Three separate sets of conservation scores with different comparative species were extracted from UCSC genome browser - vertebrate, primate, or placental. Thus, features such as first_ exon_vertebrate emerge from this set [5].

There is a total of 245 features selected by CFS in the feature selection process. There are 74 features of methylation are selected, 75 features of histone, 90 features of nucleotide composition, 4 of the conservation features and rest 2 of the element length. We first studied the connection between the characteristics chosen. Using hierarchical clustering on absolute correlation values between characteristics, we discovered that the chosen characteristics tend to cluster by type of information as anticipated. The conservation characteristics in the coding regions (CDS) are grouped together, for instance, and so are most methylation characteristics Table 1. As expected, the promoter's CpG islands are very essential for gene expression prediction, as evidenced by the three chosen and extremely correlated characteristics of CG structure, TSS200 GC, TSS200 CG and TSS200 CGG.

Categorization by Data Type	**Number of selected features**
Histone	75
Methylation	74
Nucleotide Composition	90
Conservation	4
Element Length	2
Categorization By Gene Location	**Number of selected features**
TSS200	2
CDS	2

First Exon	87
Full Transcript	87
TSS1500	2
UTR5	2
First Intron	57
Last Exon	2
Last Intron	2
UTR3	2

Table 1: Feature Selection.

Feature Selection

Feature selection is done using four different methods: Correlation Feature Selection (CFS), Gain Ratio, Principle Component Analysis (PCA) and ReliefF. CFS is a well-known non-linear measure of correlations which is based on mutual information approach. In this method, redundancy is reduced and relevance is enhanced by selecting an optimal set of features. The correlation of class to the feature measured by mutual information is the relevance, while the correlation between two features is the redundancy [8]. To reduce the number of selected features, redundancy is reduced. CFS consists of a built-in system for the selection of the number of features. In the Gain Ratio algorithm, similar to CFS a decision tree is employed. In Gain Ratio, as the name suggests, a ratio of information is obtained through which the bias of Information Gain (favoring features with a large number of data) is eliminated. Gain Ratio is a ranker system therefore, every ration has a matched ranked output. In PCA, the number of features is reduced by constructing a smaller number variable that has a sufficient amount of information obtained from the original features. PCA finds the eigenvectors of the covariance matrix with the highest values and further uses that data to project it into a new subspace [www.chrisalbon.com]. In other words, PCA converts a matrix of n features into a dataset that has less than n features.

ReliefF is the advanced and improved version of Relief. It works on hit and miss algorithm. The weight factor is constantly upgraded by using the Manhattan distance of hit and miss from any random instance. This weight factor is further used to calculate the score of relevance. A threshold value of relevance is finalized and the features that fall above this range are considered to be 'selected'. The Relief was not effective on partial data and it was performed on the nearest hit and miss. These two drawbacks are rectified in the ReliefF method as it can handle the incomplete data and operate on the average of k-near hits and misses. Just like Gain Ratio, ReliefF also works as a ranker system, consisting of matching ranked output for every input.

Classifiers

Classifiers are used for high dimensional feature spaces without mapping the points into the spaces. The data sets and their features are classified and categorized using the following classifiers. Classifiers are thus a set of mathematical expressions that are implemented by a predefined set of algorithms and result in mapping the input data. In machine learning, the computer learns the patterns from input entry and uses the learned pattern to classify the new data. Under the umbrella of machine learning, classifiers enhance data mining and data processing results.

a) **SVM:** Support Vector Machines are the classifiers that classify the data based on the planar theory. They focus on the decision planes that define the decision boundaries. Planar mathematics then decides the output of the process.

b) **Gaussian SVM:** The Support Vector Machine classifier having Gaussian kernel is a weighted linear combination of kernel function measured between a single point of data and individual support vectors.

c) **Linear SVM:** Linear SVM is a fast data mining algorithm for dealing with multiclass classification problems out of an extremely large set of data. The set of mathematical functions called kernel is applied to accomplish the process.

d) **Logistic Regression**: This classification method is used to predict the probability of an instance that belongs to its default class. It can be either 0 or 1. It is a linear algorithm that has non-linear transform on its output.

e) **Naïve Bayes Classification:** Naïve Bayes classifiers are the probabilistic classifiers that work on the principle of Bayes' Theorem. The algorithm has independent assumptions among the features.

f) **Random Forest Classifiers:** Random Forests are used for classification and regression by constructing decision trees during training. It thus makes the decision tree based on the subsets that are selected randomly from the dataset.

g) **K-Nearest Neighbour Classifier:** It is a non-parametric model of classification, which can also be used for regression. It stores all the possible cases and uses them to classify the new cases based on the similarity index.

h) **Neural Network:** This method is a form of optimization of the network. This algorithm resembles the artificial neural structure of the human brain and is used for recognizing particular patterns in the datasets.

Results

Four distinct types of data were used to extract the features of cancerous cells and four different feature selection methods were used for distinguishing the features. The percentage ratio for feature selection is raw, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, and 50%. Further, the results are concluded using eight different classifiers. 10-fold cross validations are done for different training and testing ratios. (90:10, 80:20, 70:30, 60:40). After analyzing the results of data processing by CFS feature selection method, the cells showing the best result are extracted and mentioned in Table 2. The results show the highest AUC of 0.999 is obtained by Logistic Regression classifier when 60% features are selected for a

training-to-testing ratio of 60:40. The highest accuracy of 100% is obtained by SVM, Linear SVM, Naïve Bayes, Logistic Regression and Random Forest classifiers for different percentage of features selected and training-to-testing ratios. Random Forest shows to be the most efficient classifier achieving 100% accuracy for a range of selected features.

CFS				
Ratio	Feature selection % ratio	Classifier	Accuracy	**AUC**
60:40	60%	Logistic Regression		0.999
80:20	80%, 75%	SVM	100%	
70:30	85%	Linear SVM	100%	
90:10	75%	Linear SVM	100%	
60:40	75%	Naive Bayes	100%	
70:30	70%	Logistic Regression	100%	
80:20	95%, 65%, 60%	Random Forest	100%	
90:10	85%	Random Forest	100%	

Table 2: CFS Results.

Table 3 indicates the analysis of Gain Ratio feature selection method, the cells showing the best result are extracted and mentioned. The results show the highest AUC of 1.0 is achieved by Naïve Bayes and Random Forest classifiers for 75% and 65% features selected, respectively. The highest accuracy of 100% is achieved by Gaussian SVM, Linear SVM, Naïve Bayes and KNN classifiers for different percentage of features selected and training-to-testing ratios.

Gain Ratio				
Ratio	Feature selection % ratio	Classifier	Accuracy	AUC
60:40	75%	Naive Bayes		1.0
80:20	65%	Random Forest		1.0
80:20	Raw	Gaussian SVM	100%	
90:10	85%	Linear SVM	100%	
80:20	70%	Linear SVM	100%	
70:30	60%	Linear SVM	100%	
70:30	60%	Naive Bayes	100%	
70:30	60%	KNN	100%	
80:20	50%	KNN	100%	
90:10	85%	KNN	100%	

Table 3: Gain Ratio Results.

Table 4 shows the results of Principal Component Analysis (PCA) with different classifiers. The highest AUC value of 0.999 is achieved by Random Forest, KNN, Neural Networks, Linear SVM and Naïve Bayes classifiers for different percentage of selected features and training-to-testing ratios. The highest accuracy of 100% was achieved by KNN, Linear SVM, SVM, Logistic Regression and Neural Networks classifiers for different percentage of features selected and training-to-testing ratios.

PCA							
Ratio	Feature selection % ratio	Classifier	Accuracy	Ratio	Feature selection % ratio	Classifier	**AUC**
80:20	75%	KNN	100%	70:30	85%	Random Forest	0.999
80:20	65%	Linear SVM	100%	90:10	85%	KNN	0.999
70:30	55%	Linear SVM	100%	70:30	85%	Neural Network	0.999
60:40	65%	SVM	100%	90:10	80%	Linear SVM	0.999
70:30	60%	Logistic Regression	100%	90:10	55%	Naive Bayes	0.999
80:20	55%	Neural Network	100%				

Table 4: PCA Result.

Table 5 shows the analysis of ReliefF feature selection method for the best results. The results show that the highest AUC of 1.0 was achieved by SVM for 90% features selected for training-to-testing ratio of 80:20. The best accuracy of 100% was achieved by SVM, Gaussian SVM, Linear SVM, Logistic Regression and Random Forest classifiers for different percentage of features selected and training-to-testing ratios.

ReliefF				
Ratio	Feature selection % ratio	Classifier	Accuracy	AUC
80:20	90%	SVM		1.0
80:20	60%	SVM	100%	
90:10	95%, 75%	Gaussian SVM	100%	
90:10	95%	Linear SVM	100%	
60:40	65%	Linear SVM	100%	
80:20	95%	Logistic Regression	100%	
60:40	85%	Logistic Regression	100%	
90:10	95%, 80%	Random Forest	100%	

Table 5: Relieff Results.

Discussion

The model consists of a large number of gene data points in the training set as well as the testing set. Evaluation is based on four distinct feature selection methods. Eight different classification methods, linear as well as non-linear methods were applied. A total 245 different features are selected with the best feature selection method (CFS method) spanning the methylation, histone, human genome as well as CHIP-Seq data. Initially, the relationship between these four is derived. Based on that, the clustering of selected data is obtained. In the entire process, CpG islands are proved to be important for the gene expression prediction. Also, a collinear relation is observed between methylation and histone as some of the methylation features are found to be clustering with histone modification features [7]. The entire data is classified based on eight different classifiers namely, Support Vector Machine (SVM), Gaussian SVM, Linear SVM, Naïve Bayes, Logistic Regression, Random Forest, K-Nearest Neighbors (KNN) and Neural Network.

Though various approaches figure out the modification in epigenetic as well as genetic expression, a well-structured quantitative approach that highlights the accurate prediction of up and down-regulation of gene expression is still lacking. It has been noted quite often that reliable epigenetic data is obtained but genetic data is missing. Epigenetics measurement is possible in several data sets for which genetic quantification is difficult. In such cases, a predictive method can effectively provide the required information. Apart from providing the prediction of gene expression, this model also expresses the relative importance of genome data and their genomic location. Certainly, CpG methylation data consists of more predictive values for various genetic expressions. Although various histone modification data can be used for the study, they are quite expensive as compared to CpG methylation. All the parameters in this model are useful in extracting some sort of information at the genetic level. Many features obtained from methylation and histone modification are based on the annotations from Illumina 450K for DNA methylation.

Figure 1 is the graphical view of CFS feature selection method with the different classifiers. The figure shows 100% accuracy for different classifiers with varying percentage of selected features and training-to-testing ratios. Random Forest gives the best performance with different percentage of features selected. SVM, Linear SVM, Naïve Bayes and Logistic Regression classifiers also show highest accuracy with CFS feature selection method. In order to avoid overfitting, we have to add a penalty against unpredictability to decrease the level of overfitting or fluctuation of a model by including more bias, and by creating trees on random subsets. We have implemented 10-fold cross validation which also plays an important role in avoiding overfitting.

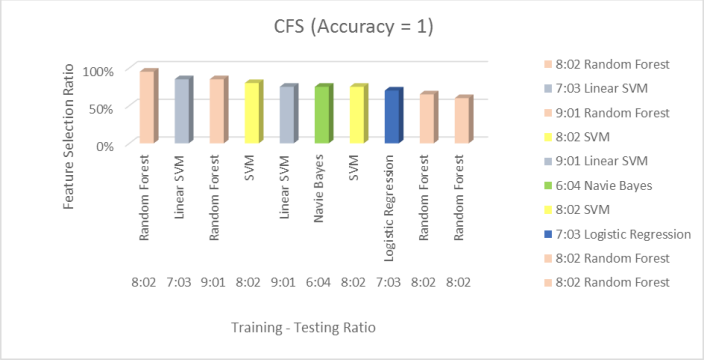

Figure 1: Graphical Representation of CFS Output Data.

For analytical study of the performance of Gain Ratio feature selection method, a graph has been plotted considering the data for different classifiers when accuracy is 100%. Figure 2 represents the output graph taken with data at different training-testing ratios. In most of the cases, Linear SVM and KNN give the best performance with different percentage of features selected. Naïve Bayes also achieved the highest accuracy.

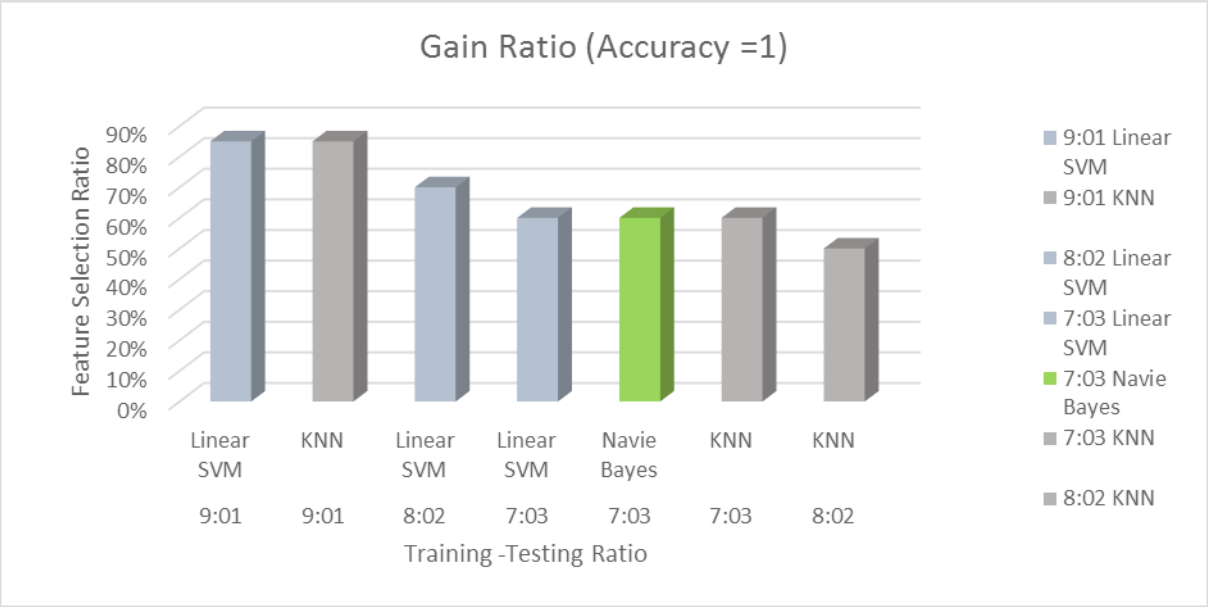

Figure 2: Gain Ratio Output Data Graph.

Figure 3 displays the comprehensive graph for the performance of PCA with different classifiers. Five classifiers – SVM, Linear SVM, KNN, Logistic Regression and Neural Network, achieve 100% accuracy with different percentage of features selected.

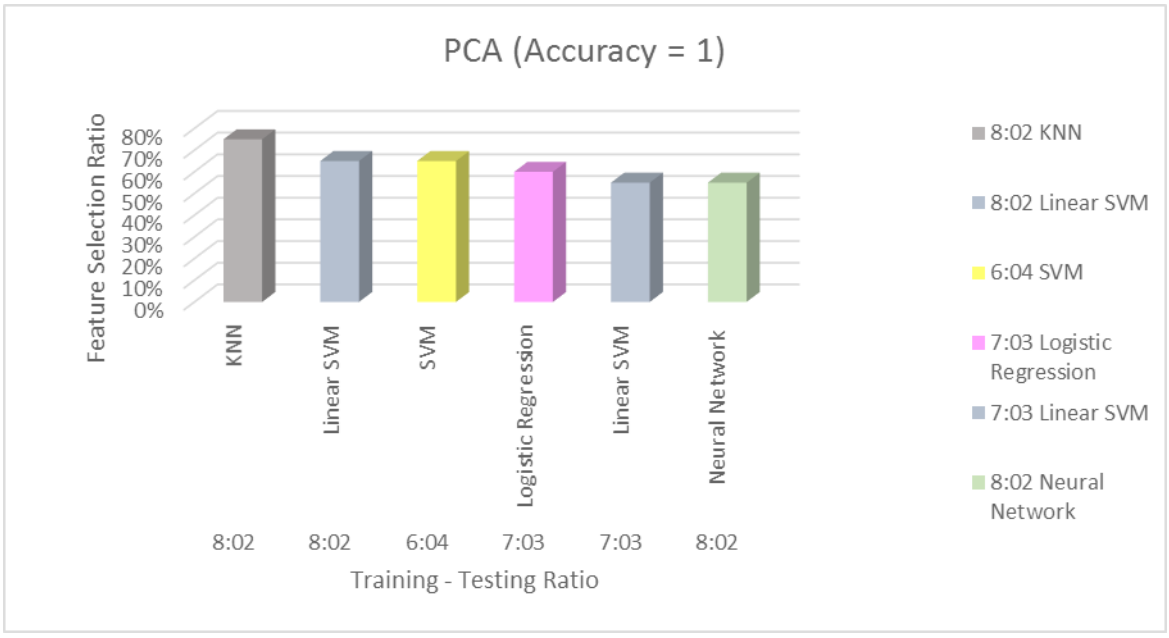

Figure 3: Output Graph for PCA.

Figure 4 represents the graph for ReliefF feature selection with different classifiers achieving 100% accuracy. Five classifiers – SVM, Gaussian SVM, Linear SVM, Logistic Regression and Random Forest gives the best result with different percentage of features selected.

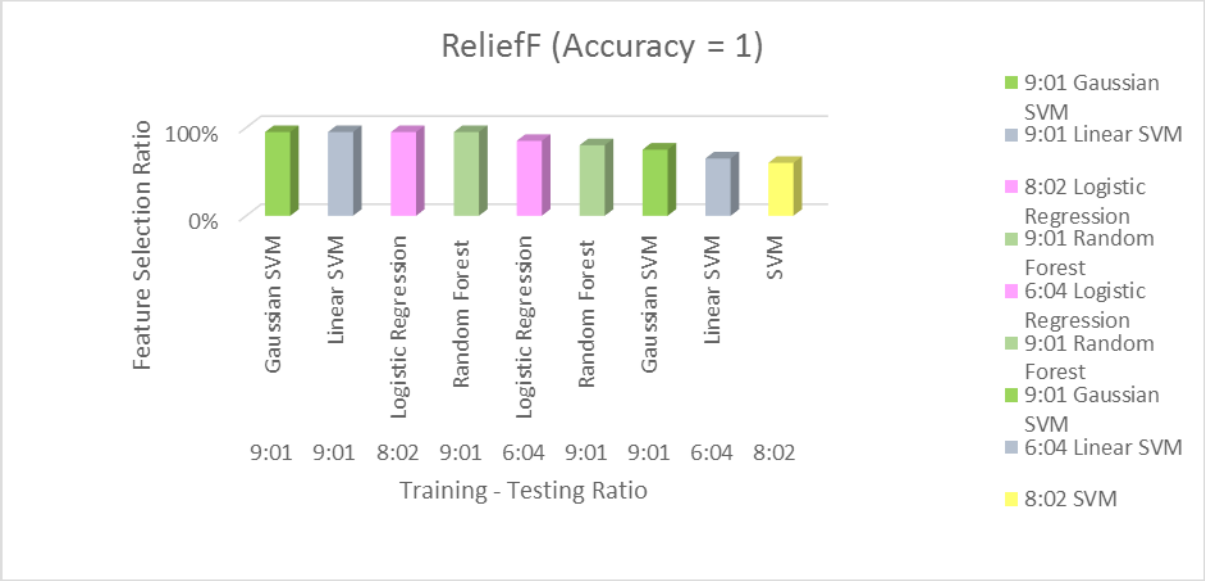

Figure 4: ReliefF Graph.

Limitations and Future Discussion

The current model lacks in covering all the histone data available and rather covers up only H3k4me3, H3k27me3, H3k36me3, and H3k4me3 histone marker CHIP-Seq data. The model accuracy is affected by the heterogeneous nature of the data. However, as the amount of data included increases, the accuracy of the model also improves. In the presented model, the dataset is split into training and testing sets and both of them are individual sets. However, if methylation data and paired RNA-Seq can be identified, the model can be replicated with attractive accuracy rates. This model is limited to the breast cancer data and is not tested on the database of any other type of cancer or any other epigenetic disease. Subsequently, it is quite evident from the available data that the epigenetic prediction is quite complex and includes the consideration of a wide number of parameters. The model has limitations in terms of the number of datasets and the selection of features out of them. Accuracy and reliability can be certainly enhanced if all the affecting parameters can be included in a single model.

Conclusion

An approach based on epigenetic data is presented that can predict various gene expression in case of breast cancer. It is evident that CpG methylation data is the key to unlock all the necessary data for different predictions. Additionally, promoter regions hold a key position in supporting the outcome of the entire process. The best model selects 245 features from 1564 features. According to the best feature selection method CFS, Random Forest classifier selects the best model for different characteristics achieving 100% accuracy the greatest number of times. Overall, CFS feature selection method selects the best model and Linear SVM classifier gives the best performance in predicting the breast cancer.

Author contributions

All authors contributed equally in brainstorming and writing the manuscript. NP performed the experiments, conducted bioinformatics data analysis and wrote the manuscript. CJ conceived and designed bioinformatics analysis, reviewed the manuscript. KP designed the experiments, conducted bioinformatics data analysis, reviewed and revised the manuscript.

Conflict of Interest Statement

The authors declare no conflict of interests.

References

1. Bock C, Lengauer T (2008) Computational epigenetics. Bioinformatics 24: 1-10.
2. Daura-Oller E, Cabre M, Montero MA, Paternain JL, Romeu A (2009) Specific gene hypomethylation and cancer: New insights into coding region feature trends. Bioinformation 3: 340-343.
3. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.
4. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261-282.
5. Hall MA, Smith LA (1999) Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. FLAIRS Conference: 235-239.
6. Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6: 65-70.

7. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32 Database: D493-496.

8. Laird PW (2010) Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet 11: 191-203.

9. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

10. Li J, Ching T, Huang S, Garmire LX (2015) Using epigenomics data to predict gene expression in lung cancer. BMC Bioinformatics 16: S10.

11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAM tools. Bioinformatics 25: 2078-2079.

12. Lim SJ, Tan TW, Tong JC (2010) Computational Epigenetics: the new scientific paradigm. Bioinformation 4: 331-337.

13. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change anddispersion for RNA-Seq data with DESeq2. Genome Biol 15: 550.

14. Pages H, Aboyoun P, Gentleman R, DebRoy S (2009) String objects representing biological sequences, and matching algorithms. R package version 2.

15. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28: 1057-1068.

16. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

17. Quinlan AR, Hall IM (2010) BED Tools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

18. Rhee JK, Kim K, Chae H, Evans J, Yan P, et al. (2013) Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. Nucleic acids research 41: 8464-8474.

19. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserve delements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

20. Smyth GK (2005) Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor: 397-420.

21. Tabl AA, Alkhateeb A, Pham HQ, Rueda L, ElMaraghy W, Ngom A (2018) A Novel Approach for Identifying Relevant Genes for Breast Cancer Survivability on Specific Therapies. Evolutionary Bioinformatics 14: 1-9.

22. Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A (2019) A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Front Genet 10: 256.

23. Wild L, Flanagan JM (2010) Genome-wide hypomethylation in cancer may be a passive consequence of transformation. Biochimica et biophysica acta 1806: 50-57.