GPELM: AN INTEGRATIVE ANALYSIS FOR BREAST CANCER SURVIVAL PREDICTION

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Arwinder Dhillon

(Roll No. 801732006)

Under the supervision of:

Dr. Ashima Singh

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

PATIALA – 147004

June 2019

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*GPELM: AN INTEGRATIVE ANALYSIS FOR BREAST CANCER SURVIVAL PREDICTION* ", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Ashima Singh* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Arwinder Dhillon

Signature:

(Arwinder Dhillon)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Ashima Singh)

Assistant Professor,

CSED, TIET, Patiala

Acknowledgement

I would like to thank God for blessing me with all the strength and resources required to complete this task. I would like to express my gratitude to **Dr. Ashima Singh**, Assistant Professor, Computer Science & Engineering Department for guiding me through the whole process and providing me with your knowledge and experience.

I am also heartily thankful to **Dr. Maninder Singh**, Professor and Head, Computer Science & Engineering Department for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection.

I would like to thank my friend and Phd scholar **Ms. Parampreet Kaur**, for being there for me always. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Arwinder Dhillon

Arwinder Dhillon

(801732006)

Abstract

High-dimensional datasets comprising genomic data, proteomic data, pathological images data and so on have taken noticeable toll now-days in healthcare research. Handling these datasets require appropriate knowledge about tools and techniques for efficient prediction like disease detection, survivability analysis, and biomarker identification etc. Researchers are working hard to achieve high accuracy in such predictions but the performance achieved in predictions for breast cancer patients is not sufficient due to high risk diseases like cancer. In this thesis, we proposed a framework called Genomic Pathological Extreme Learning Machine (GPELM), which is an invariant of Extreme Learning Machine (ELM). This package comprises of ensembling of six different models. These models consist ELM with Buckley James estimator (ELMBJ), Ensemble of ELMBJ (ELMBJEN), ELM with regularized Cox model (ELMCOX), Ensemble of ELMCOX (ELMCOXEN), ELM with gradient based boosting (ELMBOOST) and ELM with likelihood-based boosting (ELMCOXBOOST). The dataset has integrated genomic data (gene-expression, copy number alteration, DNA methylation, protein expression) and pathological images data for breast cancer survival prediction. These six models are present in the survELM package. GPELM is compared with other cox models and their related performance parameters comprising sensitivity, precision, accuracy, MCC, AUC, AUPR, hazard ratio and concordance index are calculated. GPELM has achieved 85% of accuracy for breast cancer survival prediction and there is 5% increase in each of the performance parameter taken into consideration. The purpose of this thesis is to predict the survival of breast cancer patients with best accuracy.

GPELM gives a commendable contribution in survival prediction of breast cancer patients as very little increase in performance is considered significant for breast cancer survival. All the results achieved in the present research show the usefulness of GPELM for breast cancer survival prediction. GPELM package is also implemented and tested on lung cancer data. Therefore, the package can also be implemented on more disease datasets having simple clinical, genomic, images and integrated datasets.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Computational Bioinformatics is one of the highest researched topics in Healthcare. It is the recognition of some particular diseases and providing preventive actions to improve the health of a person. Different researchers are working on healthcare for early detection and preventions of one of the deadliest diseases like cancer. The human body consists of cells. Cancer occurs when any cell of the body grows abnormally. The collection of these abnormal cells leads to tumors. From this collection, a cell detaches itself from other cells and move to other body parts through blood vessels and leads to the tumor in various parts of the body. This cell movement is called metastasis [1]. The complete structure of how cancer occurs is shown in figure 1.1.

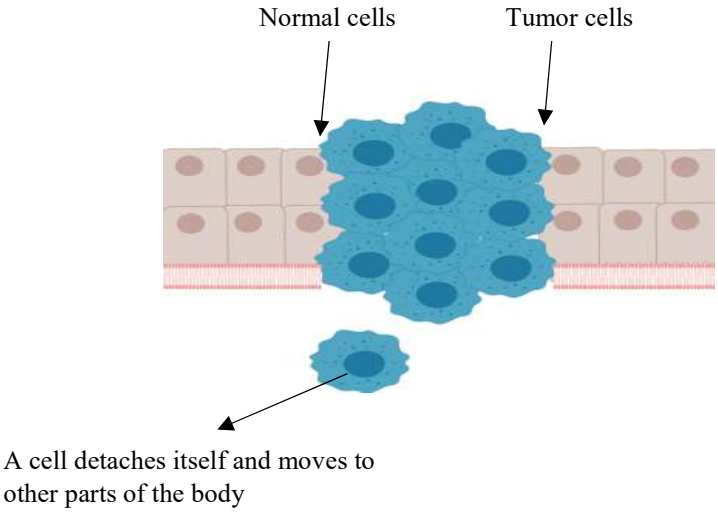

Figure 1.1 Cancer Cell

According to the National Cancer Institute (NCI), around a hundred types of cancer are present including breast, prostate, ovarian, liver, bladder cancer and so on. The mortality rate for patients dying from different cancer is 1,50000 for lung cancer, 50,000 for colon cancer and, 46000 for breast cancer which is the third most dangerous disease. NCI shows that breast cancer is the most common type of cancer in the US and needs to be diagnosed early.

1.1 Breast Cancer

Breast Cancer is a common disease in both gender but mostly found in women. It is the ruling source of death in females. In India, the survival rate of breast tumor patients is very low having 66.1% only. The death rate of patients dying from malignancy is 11.6 per 10,000 women [2]. As malignancy occurs due to unusual growth of cells, there are genes present in the nucleus of the cells. These genes are responsible for the movement of cells. Sometimes anomalous changes may turn off or turn on the genes. This turns off and turns on may lead to tumor. Breast tumor either starts in the glands which produce milk known as lobules or in the ducts which transfer the milk from lobules to nipples. After the tumor may expand to the lymph nodes which act as a filter for outside particles. From the lymph nodes, it passes to other areas of the body [3]. The complete structure of the breast is explained in figure 1.2 below.

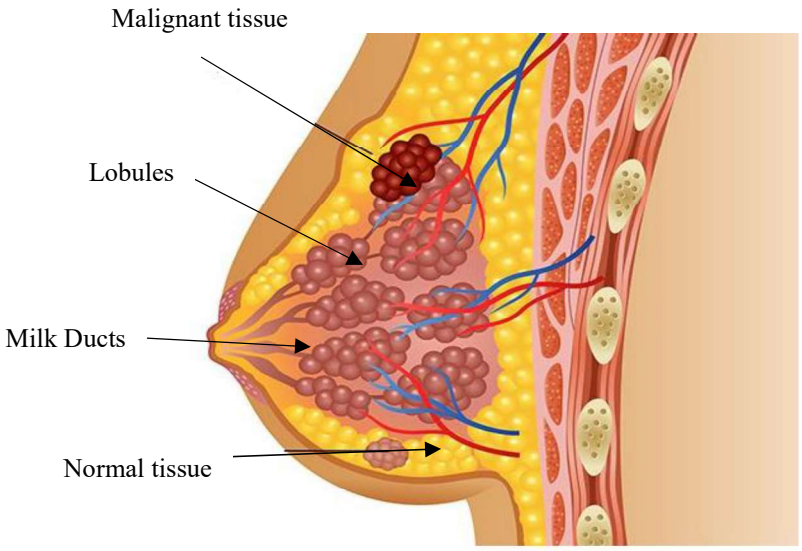

Figure 1.2 Breast Anatomy [59]

1.1.1 Types of Breast Cancer

There is various kind of nastiness available in the teat which can be either grow to other pieces of the body known as obtrusive or which does not grow to other areas known as non-obtrusive [4]. The types of malignities are explained below:

a) Intraductal Cancer (IC): IC embarked from inner of the minute tubes recognized as channels and is of non- obtrusive type.

b) Invading Ductal Cancer (IDC): IDC embarked from the channel and start growing to other stuff of the teat as of its obtrusive nature.

c) Lobular Cancer in Situ (LCIS): LCIS embarked from internal of the milk-preparing secretor available in the segments of the udder secretor and is non- obtrusive in nature.

d) Infiltrating Lobular Cancer (ILC): Like IDC, ILC embarked from the segments and extend to other regions of the udder. It is obtrusive in nature.

e) Inflammatory Udder Cancer (IUC): IUC embarked by chocking the bodily fluid vessels available in the skin and extends very fast to other regions. The udder got red and swollen in this type.

f) Metastatic Udder Cancer (MUC): MUC embarked in the udder and grows to other regions like lungs, liver and so on.

Table 1.1 Types of Malignities

Type	**Place**	**Obtrusive / Non-Obtrusive**
Intraductal Cancer	Inside the tiny tubes	Non-Obtrusive
Invading Ductal Cancer	Inside the tiny tubes	Obtrusive
Lobular Cancer in Situ	Inside the sac	Non-Obtrusive
Infiltrating Lobular Cancer	Inside the sac	Obtrusive
Inflammatory Udder Cancer	Lymph vessels	Obtrusive
Metastatic Udder Cancer	Breast	Obtrusive

1.1.2 Diagnosis of Breast Cancer

For diagnosis [5], a specific procedure or test is followed which is defined below

a) Self-Exam: In this, the udder and bodily fluid nodes are examined by the doctor to see any chunk or unusual tissue.
b) Mammogram: An X-Ray of the breast is performed to screen cancer. If any lump is found then the patient is advised for further evaluation.
c) Ultrasound: If any lump is found then to check whether this lump is a solid or watery cyst, a deep structured image is captured with the help of sound waves.

d) Biopsy: In this, a sample of tissue from the doubtful area is taken with the help of a needle to evaluate cancer. If cancer is found then it is further evaluated to check the type of malignity.

1.1.3 Treatment of Breast Cancer

There are many ways a malignity can be treated resting upon the magnitude and level of cancer which is explained below:

a) Surgery: In this resting on the magnitude of malignity, either part of stuff along with the whole sum stuff is abolished or a complete udder is abolished if the malignity is large. If a teat is abolished then it is recreated by taking a part from other body stuff. It consists of breast-conserving surgery in which an area of malignity is removed, mastectomy where the whole udder is abolished, recreation and, a biopsy is performed to check whether cancer has spread to sentinel bodily fluid node known as bodily fluid node surgery.

b) Radiotherapy: In this, radiations are used to kill the remaining malignancy. It is performed for 3-5 days in a week and can be extended for 6-7 weeks. It consists of breast radiotherapy in which a part of malignity is removed to abolish the remaining malignant cells, chest wall radiotherapy in which the whole breast is abolished performed at the chest wall to remove remaining stuff and, breast boost which is performed to kill malignant cells which are available in the axilla of bodily fluid nodes.

c) Chemotherapy: In this, the left cells are terminated with the help of cytotoxic which is an anti-cancer medicine. If the malignity got enlarged and spread to bodily fluid nodes then it is not possible to abolish cells with chemotherapy. There it is used only to decrease the size of malignancy. It is of two types, adjuvant which is performed after surgery and neo-adjuvant which is performed before surgery.

d) Hormone Therapy: There are some malignity cells which origins due to hormones present inside the body. These are called a positive hormone receptor. Hormones are stopped from expanding with the help of hormone therapy. Hormone therapy is performed before or after the surgery depending on the context of the tumor. It is performed before surgery in order to shrink the tumor so that its size gets reduced and it is easy to abolish it. Hormone therapy treatment is used only for the treatment of breast cancer. The complete treatment process is shown in figure 1.3 below:

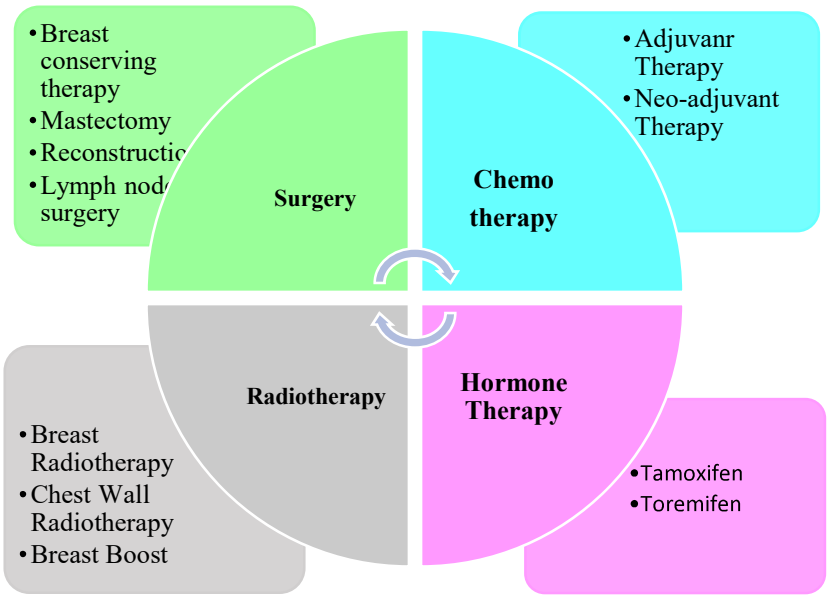

Figure 1.3 Treatment of breast cancer

1.2 Types of Dataset

Different types of data have come into view in healthcare nowadays including clinical data, sensor data, omics data and so on. This type of data includes different mining methods to extract the more relevant features and the different algorithms need to be trained for better future prediction. For prediction of breast cancer, genomic and pathological images dataset comes into view which is described below and is diagrammatically shown in figure 1.4 below

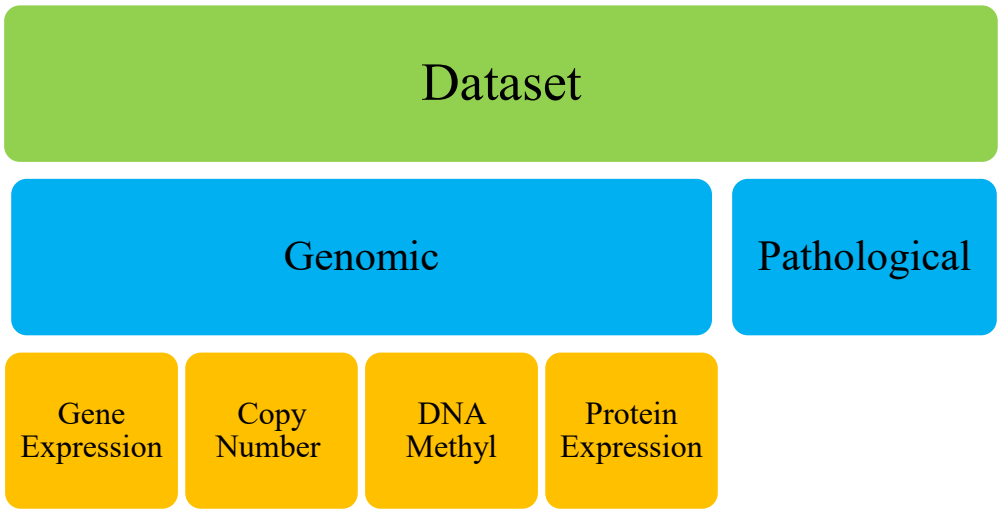

Figure 1.4 Types of Dataset

1.2.1 Genomic Data

Genomic data is collection of Gene Expression, Copy Number Variation, Protein Expression, and DNA Methylation data which consists of genes values and are described below. The data can be retrieved from the TCGA portal. This data is initially in raw form which is difficult to understand by normal people. So, to work with, data need to be preprocessed in normal form.

a). Gene Expression: Gene Expression data defines how the information is generated from the gene and how to use that information for making a useful gene product. The human body is made of a cell. Each cell contains miRNA data which is used to produce information. The flow of information starts with DNA. From there it moves to RNA and then to protein. this conversion of DNA to RNA is known as transcriptome data which is used in our research. In every second, thousands of transcripts are produced by each cell. These transcripts are responsible for affecting the activity of the body.

b) Copy Number Alteration: Copy Number Alteration data is a very important component of genome data and is described as a DNA segment of one or more kilobase as it is a variable number compared to gene-expression. This variation leads from several insertions, deletions, and update on the chromosome number which leads to a large variety of data. These variations are implicated in somatic cells which leads to a tumor.

c) DNA Methylation: In DNA Methylation, the methyl group is converted to cytosine ring of DNA at the 5^{th} position. It is usually present in the mammalian genome. It modifies the function of genes and affects gene expression data by changing patterns of DNA. It is a vital process which is very much important for the growth of ordinary cells. In order to inactivate X-chromosomes, DNA methylation plays an important role.

d) Protein Expression: Similarly, Protein Expression data defines how the proteins are modified in the cells of a human body. Blueprints of proteins are stored in DNA and further decoded to produce RNA. RNA produces information in the form of proteins. Protein expression is basically finding the proteins present in a sample of tissue which can be obtained at a particular time based on some conditions. When all the proteins are obtained, we then select the one which changes most. These changed proteins are the one which leads to cancer.

1.2.2 Pathological Images

Pathological images are whole slide images of cell, tissue or body fluid to detect cancer. In this, a glass slide is converted into a digital image which can be managed or analyzed on a computer screen. Tissue slide of the affected area is taken which can be seen in a microscopic environment in order to get the information. We can also generate numeric values of the area using various algorithms. With the help of pathological images, doctors are able to find the affected area easily whose surgery of biopsy is needed.

1.3 Machine Learning

Handling healthcare data comprising electronic health record (EHR), omics data, clinical data and so on is a tedious task for a normal human being. The early detection and prognosis of some disease become a necessity. This can be done easily with the help of Machine Learning. Machine learning is a learning process which is basically programming the computers based on the historical data and from previous experience. Sometimes we can do work easily what we cannot explain that how can we do that. In that case, learning is required [6].

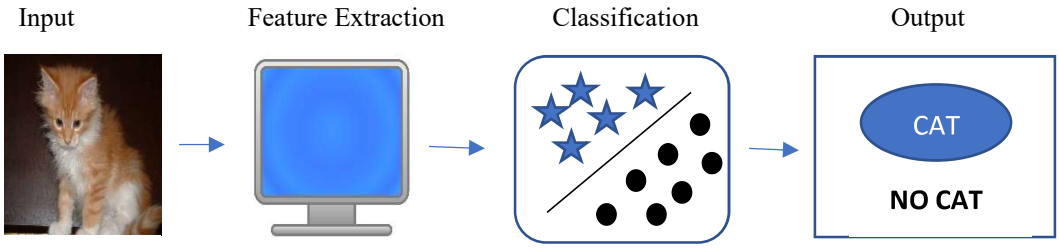

Figure 1.5 Machine Learning [60]

Machine learning is an act of learning and improving automatically from past experience without any detailed programming. It is used to look for patterns in the data and use these patterns for training. Once training is finished or once computers learn from the data then we will use this training as testing for new data items and check for the outcome as shown in figure 1.5 above. The purpose is to clearly understand the data and use that data for modeling so that it can be easily grasped by humans.

1.3.1 Machine Learning Algorithms

There are 4 machine learning algorithms comprising supervised learning, unsupervised learning, semi-supervised and reinforcement learning and are discussed below:

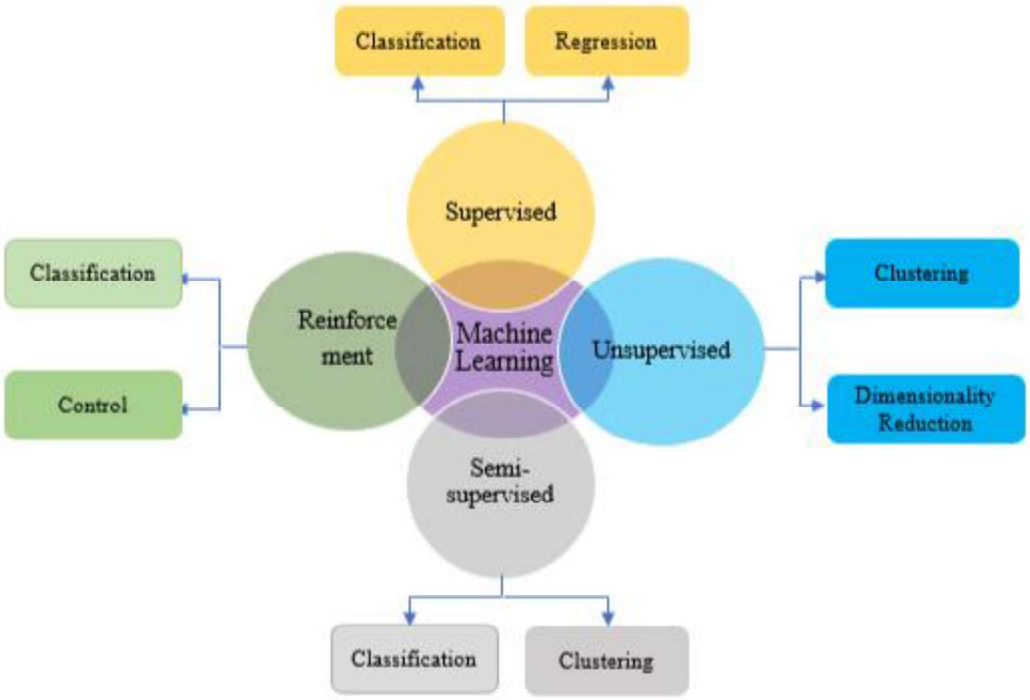

Figure 1.6 Types of Machine Learning Algorithms

a) Supervised Learning: Supervised learning involves training the model on the labeled data and uses this trained model to make predictions on the new data. It involves splitting of data into two sets including the training set and testing set. First, the model is trained on a training set and afterward, the performance is tested on the testing set. The performance of the model can be evaluated using performance metrics. Supervised learning can be a classification problem or regression problem. In supervised classification, the labeled value is a discrete value.

b) Unsupervised Learning: Unsupervised Learning also involves training of the data except for the fact that the labeled value or target value is not known. In this, the machine tries to cluster the similar type of data by finding the hidden pattern. Rather than making a prediction, the main aim of unsupervised learning is to discover the patterns. The performance of the model in unsupervised learning cannot be evaluated as the label value is absent or unknown. The algorithms involved in unsupervised learning are K-mean clustering, Association Rule Mining, Topic Modeling, and Dimensionality Reduction Techniques.

c) Semi-supervised Learning: As supervised learning works on labeled data and unsupervised learning on unlabeled data, then a lot of information is lost from labeled data which can be obtained from unlabeled data. So, in this case, semi-supervised learning comes to mind. It is a mixture of supervised and unsupervised learning in which it takes both the unlabeled and labeled data. Labeled data should be of shorter length as compared to unlabeled data. The idea behind semi-supervised learning is that there is a considerable change in performance when both labeled and unlabeled data is used in conjunction. The training set used is of shorter length. It is normally used to detect outliers.

d) Reinforcement Learning: Reinforcement Learning works by developing a system which improves its performance by taking feedback from the environment and taking possible steps to improve them. It is an act of learning from the environment by interacting with it without any help from humans. It is an iterative process.

The different types of machine learning algorithms and their applications are shown in figure 1.6 above.

1.3.2 Extreme Learning Machine

Extreme Learning Machine works on single-hidden layer feed forward neural network in which input weight value and hidden nodes value is chosen randomly and output is generated accordingly. These values need not be inherited from the parent's node. From past years, ELM has been used for prediction on high dimensional datasets for survival analysis. It can be applied to classification, regression, and clustering problem for prediction purpose. Normally it is difficult for normal neural networks to train models because of a large number of classes [7,8]. Also, it is a very slow process and takes a lot of time to train. So, for fast and accurate training, ELM is a perfect choice because the only cycle is required to train the model. The ELM model is combined with the given models in the proposed approach.

1.3.3 Buckley James Estimator

Buckley-James estimator was introduced in 1979 by Buckley and James which make use of least square estimation method for censored data. To handle censoring, the update is required for handling data based on non-parametric equations [9]. This update is required to group various forms of censored data. It is a regression technique and also called an

accelerated failure time model. It is best suited for simulation problems and is better than regular cox models.

1.3.4 Regularized Cox Models

Regularized cox models are used to identify the effect of some variable when some event occurs on survival time. The variables used for prediction have a constant effect on survival time, so they are non-parametric in nature. These provide better results than normal kalpen-meier curves. They are used to calculate hazard ration based on survival time and censored values.

1.3.5 Gradient and Likelihood-Based Boosting

Both gradient boosting and likelihood boosting are an ensemble of decision tree for accurate predictions. It works by providing optimization of differential loss function and steps followed are the same as boosting. It is a type of black box learning because it works on libraries which are already present without knowing the details of how it is working. Normal difference between decision and ensemble is that ensemble reduces noise and bias values. Gradient and likelihood boosting works by learning from previous mistakes [10]. When one prediction is made based on a decision tree, random forest and so on, new prediction takes less time because the previous already reduces the noise values and are not included in the new prediction. So, it works by learning on the error value. It belongs to a classification and regression problem.

1.4 Performance Parameters for Survival Analysis

The parameters which we use to achieve the performance in our model for survival are described below

1.4.1 Sensitivity

Sensitivity is defined as the degree of total true positive or positive cases which are predicted as true. It is also known as recall. It is given as the ratio of True Positive (TP) with the sum of True Positive (TP) and False Negative (FN). True Positive means patient having cancer is actually suffering from cancer and False Negative means person having no cancer are predicted as having cancer.

$$Sensitivity = \frac{TP}{TP+FN} \quad (1.1)$$

1.4.2 Specificity

Specificity is defined as the degree of actual negative value and is predictive as negative. It is the ratio of True Negative with the sum of True Negative (TN) and False Positive (FP). True Negative means person not suffering from cancer has no cancer in actual. False Positive means the person predicted as having cancer has actual no cancer.

$$Specificity = \frac{TN}{TN+FP} \quad (1.2)$$

1.4.3 Precision

Precision is defined as the percentage of actual results which are true or relevant. It is given by the ratio of True Positive (TP) with the sum of True Positive (TP) and False Positive (FP) and is defined in eq. 1.3 below.

$$Precision = \frac{TP}{TP+FP} \quad (1.3)$$

1.4.4 Accuracy

Accuracy is defined as the percentage of Predicted results which are predicted correctly with respect to actual values and is given by following eq 13.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1.4)$$

1.4.5 Matthews Correlation Coefficient

It is defined as the measure of the quality of binary classification and even is used were classes of different sizes. Its value lies between -1 to 1.

-1 defines false prediction, 0 defines random and 1 defines correct prediction.

$$Matthews\ Correlation\ Coefficient = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TN+FP)\times(TN+FN)\times(TP+FP)}} \quad (1.5)$$

Here TP stands for True Positive, FP for False Positive, TN for True Negative and FN for False Negative. These can be obtained by using a confusion matrix which is a matrix between actual values and predicted values.

1.4.6 Concordance Index

Concordance index is defined as a ranking variable whose value lies with 0 to 1 where 0 represents the worst value and 1 represents the best value. More the value of concordance index, the more the better the performance is. The value of the concordance index is given by:

$$CI = \frac{1}{n} \sum_{i\epsilon\{1..n|\delta_i=1\}} \sum_{s_j>s_i} I\left[X_i\hat{\beta} > X_j\hat{\beta}\right] \quad (1.6)$$

Here n is total no of comparison, I represent the indicator and s represents the actual result.

1.5.7 Hazard Ratio

The hazard ratio is the ratio of an event occurring in one group as compared to an event in another group. The event in one group means the treatment process and event in another group means the control process. Hazard ratio value should always be less than 1. When the value is close to 0, it means that there is approx 100% reduction in risk in certain disease and value close to 1 means there is 0 % risk reduction.

1.5.8 Area under Precision-Recall (AUPR)

The area under Precision Recall shows the relation between precision and recall at different threshold values. A high AUPR means both precision and recall values are high. Higher the value of AUPR, better the model is.

1.5 Chapterization

The rest of the thesis is organized as given below:

Chapter 2 gives a detailed view of related work done based on breast cancer survival patients using machine learning algorithms.

Chapter 3 define research gaps collected from literature survey and problem statement of the work. The methodology adopted is also discussed in this section.

Chapter 4 describes the proposed integrated genomic and pathological images for breast cancer survival using extreme machine learning (GPELM).

Chapter 5 describes the implementation and result obtained by applying a machine learning algorithm for survival prediction.

Chapter 6 defines the conclusion and future scope of the research work. After that references are given.

1.6 Summary

This chapter provides detailed information about breast cancer, diagnosis, and treatment of breast cancer, types of data required for prediction, machine learning, machine learning types, and different models used.

Chapter 2

Literature Survey

In this chapter, the analysis of Breast Cancer, different types of the dataset used for prediction, machine learning, and different methods for integration of data is performed. Work done by various researchers to perform the analysis is shown as follows.

2.1 Breast Cancer

Breast cancer is a disease which evolves in cells of the body due to the mutations called changes occurs in genes. It starts from channel and lobules present in the breast that produces milk and the channel through which milk transfers. From here the cells embarked to other regions of the body that is to bodily fluid nodes under arms and move further leading to lung, ovarian, cervical, and colon cancer if not controlled in the beginning. There are different types and stages of breast cancer which are explained in chapter 1. According to WHO, it is the second type of cancer which normally present in women. It can be classified into two types benign and malignant tumor. A benign tumor means there is no cancer that is cells work normally. On the other side, malignant means there is cancer that is cells did not work accordingly. It is very difficult to detect it at a later stage. So, computers are trained to detect it at the early stages in order to reduce the death rate of cancer patients.

2.1.1 Causes of Breast Cancer

i) Hereditary

One of the main risk factors is heredity that there are chances for patients if their close family member had cancer. There are two types of genes present in the human body called BRCA1 and BRCA2 which transfers to their children which leads to cancer. There is one more gene called TP53 which also causes cancer. So, regular checkups need to be done in those cases.

ii) History of lumps

If a woman had cancer at an early age, then there are chances that she again got cancer due to abnormal growth in benign cells.

iii) Estrogen Exposure

When there is an increase in estrogen level, then there are chances of breast cancer. Estrogen is produced when the periods start until menopause occurs. Some girls get periods in early age and their menopause occurs later than average age. So, they produce a great amount of estrogen which leads to exposure and can lead to cancer.

iv) Obesity

Women who are fat or become fat after menopause due to higher estrogen level or high intake of sugar can have higher chances of breast cancer.

v) Age

Age is also one factor in which women who are older have a higher chance of cancer than the one who is younger.

2.1.2 Symptoms of Breast Cancer

The main symptom of breast cancer is the presence of the chunk in the breast or under the armpit or stiffening of the breast. Other Symptoms are:

- Continuous pain in the breast which does not even stop over time.
- Redness of skin near the breast.
- Ejection of liquid or blood from the breast.
- Removal of skin from the breast.
- Change in shape or size of the breast.

2.2 Machine Learning

Machine Learning was originated by Samuel in 1950 to play strategic games like chess. It is the mechanism of making machines to learn automatically without being explicitly programmed. The main focus of Machine Learning is to develop a computer program which can access the data and use this data for learning purpose. It is the ability of a machine to make use of statistical techniques and advanced algorithms to make a more powerful prediction and making the data-driven system more powerful by replacing the rule-based system [11]. The main component of machine learning is data which is the backbone for any model. The more relevant data is the more accurate predictions are. After

data, we need to select the algorithm based on the problem for more accurate predictions. Machine Learning can be used in many fields such as finance, retail, health care, and social data [12].

2.2.1 Supervised Learning

Any learning can have some set of features. These features are either continuous or categorical. Sometimes they can be binary also. Along with that if the features have known labels then the learning is classified as supervised learning [13]. It can be a classification problem that we can classify the problem into classes or it can be a regression problem [14]. There are many algorithms which come under supervised learning which includes Random forest (RF), Support Vector Machine (SVM), Decision tree (DT), Bayesian networks and so on.

Examples of supervised learning are

i) Classification of a certain disease.
ii) Email classification.
iii) Voice recognition

2.2.2 Unsupervised Learning

In unsupervised learning, classes of the features are not known [11]. For example, we have a basket full of fruits. We don't know anything about them. We need to separate them. So, without any knowledge, we classify them based on their color, size, shape and so on. By doing this we obtain useful classes of objects. Algorithms used for unsupervised clustering are:

i) k-mean clustering for cluster the data.
ii) apriori algorithm for association rule mining.

2.2.3 Semi-supervised Learning

As supervised learning does not deal with unlabeled data, it is not possible to classify the objects in that case. So, for this semi-supervised approach is required [15]. It deals both with classification and clustering problem by working on both labeled and unlabeled data. Semi-supervised learning uses the following approaches to solve the problem:

i) Generative Models
ii) S3VMS
iii) Graph-Based Algorithms
iv) Multi-view Algorithms
v) Self-training

2.3 Machine Learning for Breast Cancer Prediction

2.3.1 Genomic Dataset

Genomic Data consists of gene expression, copy number alteration, DNA methylation, and protein expression dataset. Gene Expression data defines how the information is generated from the gene and how to use that information for making a useful gene product. The human body is made of a cell. Each cell contains miRNA data which is used to produce information. The flow of information starts with DNA. From there it moves to RNA and then to protein. this conversion of DNA to RNA is known as transcriptome data. In every second, thousands of transcripts are produced by each cell [16]. These transcripts are responsible for affecting the activity of the body. Copy Number Alteration data is a very important component of genome data and is described as a DNA segment of one or more kilobase as it is a variable number compared to gene-expression. This variation leads from several insertions, deletions, and update on the chromosome number which leads to a large variety of data [17]. These variations are implicated in somatic cells which leads to cancer. In DNA Methylation, the methyl group is converted to cytosine ring of DNA at the 5^{th} position [18]. It modifies the function of genes and affects gene expression data. It modifies the function of genes and affects gene expression data by changing patterns of DNA. It is a vital process which is very much important for the growth of ordinary cells. In order to inactivate X-chromosomes, DNA methylation plays an important role. Similarly, Protein Expression data defines how the proteins are modified in the cells of a human body. Blueprints of proteins are stored in DNA and further decoded to produce RNA. RNA produces information in the form of proteins. The work on genomic data for prediction of breast cancer is performed by the following authors.

Ashraf Abou Tabl et al. [19] make use of hierarchical machine learning algorithms to predict the survival of breast cancer patient after treatment of cancer with some specific

therapy. Ashraf abou and authors take a sample of 347 patients which includes clinical and genomic information to identify the class based on the genes extracted using a combination of feature selection and predictive methods. The method was trained with machine learning algorithms comprising support vector machine (SVM), Naïve Bayes classifier and Random Forest (RF) and results showed that random forest produces the best result by identifying all classes correctly with the best accuracy.

Many researchers have worked on the microarray for the identification of gene signature profiles. For example, Hongyue Dai et al. [20] take DNA microarray data for the prediction of breast cancer. Authors used a sample of 117 young patients and supervised machine learning algorithm was applied to accurately predict poor gene signature which causes cancer.

Lu Zou et al. [21], proposed unsupervised learning which consists of the integration of Principle Component Analysis (PCA) and Autoencoder Neural Network on gene expression dataset for breast cancer prediction. A total of 129,158 gene expressions profiles were taken and features were extracted with a deep learning approach. The model was trained and experimental results showed that the proposed framework works well with 85 percent accuracy.

Zhi Huang et al. [22] also proposed a machine learning framework to identify for how much time the patient is going to survive after some treatment with Cox models. Authors proposed Survival Analysis Learning with Multi-omics Neural Network on a gene-expression dataset for survival prediction. Results proved that the proposed framework works well with the best accuracy.

Chen Peng et al. [23] also worked on multi-dimensional omics data. Authors presented Machine Learning framework based on Capsule Networks on multi-omics data comprising miRNA expression and DNA methylation data sample of 4572 patients were taken and features were extracted. The model was trained with Deep Learning based Capsule Network and results proved that the proposed approach performed well as compared to different machine learning algorithms.

Cheng Fan et al. [24] uses machine learning algorithms comprising of logistic regression and Cox survival models to effectively classify breast cancer into different subtypes and to

predict survival after neoadjuvant chemotherapy. Sample of 957 breast cancer patients was taken and the model was trained.it is proved from the result that the proposed approach performed best with an accurate prediction.

Patrick Murigu Kamau Njage et al. [25] proposed machine learning algorithms for improving hazard characterization in microbial risk assessment. Because of the high dimensionality of genomics data, authors defined ML-based predictive risk modeling for risk assessment. Dataset related to DNA isolation and sequencing were collected and feature extraction was performed to extract the relevant features. Machine Learning classifiers including random forest, support vector machine, a logic boost was applied and results were evaluated. From results, it was proved that logic boost performed best with an accuracy of 75%.

Chaudhary et al. [26] defined deep learning model and six machine learning algorithms comprising random forest, support vector machine, linear discriminant analysis, prediction analysis for microarrays, recursive partitioning and regression trees and generalized boosting model for prediction of estrogen receptor status in breast cancer patients based on metabolomics data. Samples of 271 patients were taken in which 204 patients are with positive estrogen receptor and 67 with a negative receptor. K-nearest neighbor was used for normalization of data. The normalized data was trained using machine learning and deep learning algorithm. From experimental results, it was proved that deep learning algorithm performed best with an AUC value of 0.93.

J.S. Marron et al. [27] make use of clustering technique to classify breast cancer in its subtypes accurately. Samples of 115 patients were taken an experiment was performed. Results showed that the proposed approach performed well by accurately classifying breast cancer into its subtypes.

2.3.2 Pathological Dataset

There is another concept called pathological images. It is difficult from a normal person to extract information from it. As a result, it is can be included as a new research area. Pathological images are whole slide images of cell, tissue or body fluid to detect cancer. In this, a glass slide is converted into a digital image which can be managed or analyzed on a computer screen. Tissue slide of the affected area is taken which can be seen in a

microscopic environment in order to get the information. We can also generate numeric values of the area using various algorithms. With the help of pathological images, doctors are able to find the affected area easily whose surgery of biopsy is needed. Work on pathological images is done by the following authors.

Jan Ziang et al. [28] proposed convolutional neural network comprising convolutional layer, small SE Resnet nodule and fully connected layer to work on histopathological images for the classification of Breast Cancer. The SE Resnet module is a modification of squeeze and excitation module. Data was taken feature were extracted. It is cleared from the results that the proposed approach performed well than other models.

On the other side, Ce Zhang et al. [29] also worked on pathological images data for the prognosis of non-small cell lung cancer. Authors take a sample of 2186 hematoxylin and eosin stained whole slide images and applied machine algorithms to classify the patients into a long-term survivor and short-term survivors learning. Results prove that the ML algorithms produced the desired results with the best accuracy.

Depending on the size of histopathology images, Jun Xu et al. [30] proposed a framework called Stacked Sparse Autoencoder (SSAE) in order to identify the different nuclei for breast cancer survival prediction. Sliding window operation on a sample of 500 histopathology images and experiment was performed. It is proved from the result that the proposed approach performed well precision-recall curve value of 78.9%.

On the other side, Ajay et al. [31] proposed a convolutional neural network for detection of the invasive tumor on whole slide images. Sample of 400 patients was taken from different sites and the experiment was performed. Results show that the proposed approach works best with the dice coefficient value of 75.86%, the positive predictive value of 71.62% and negative predictive value of 96.77%.

Dayong Wang et al. [32] proposed deep learning framework by integrating it with human diagnostic pathological images for detection of tumor in lymph nodes. The experiment was performed individually and in integration. Results proved that the integrated approach performs best with receive operator curve value of 0.995 which is better than the individual approach having roc value of 0.925.

2.3.3 Integration of Different Datasets

In this, different datasets are integrated comprising genomic subtypes integration or genomic, pathology integration to detect the effect of machine learning algorithms for breast cancer survival predictions. Work performed by various researchers on the integration of datasets is discussed below.

Olivier Gaveart et al. [33] proposed a framework on the integration of clinical and microarray data for early prediction of breast cancer with the help of Bayesian Networks. Sample of 78 patients was taken and preprocessing was performed. After that integration of clinical and microarray data was done with full integration, decision integration, and partial integration, and model was trained. From results, it is cleared that Bayesian networks performed well with a mean area under the curve value of 0.85.

Due to the complexity of multi-dimensional data, DongDong Sun et al. [34] proposed Multimodal Deep Learning Neural Network by integrating Multidimensional data (MDNNMD) for the detection of breast cancer patients. The experiment was performed on METBRIC dataset with 1860 patients which consist of a multi-dimensional dataset comprising gene-expression, copy number variation, and clinical dataset. Feature selection was performed and MDNNMD was applied. The experimental results proved that MDNNMD performed best with 82 % accuracy as compared with SVM, RF and Logistic Regression (LR).

Working on images will also helpful in the identification of breast cancer. Hui Li et al. [35] presented a new strategy by integrating genomic and radiology images for breast cancer prognosis. Sample of 91 patients was taken and the variable selection was performed with LASSO and Logistic Regression to extract the features. The experiment was performed on individual genomic and radiology dataset and on combined genomic and radiology images dataset and the results proved that integrated genomic and radiology data performed well as compared to individual datasets.

On the other side, Fei-hung Hung and Hung-Wen Chui [36] proposed another strategy of integration gene expression and protein network for identifying subtypes of cancer correctly with the help of support vector machine (SVM) classifier. Sample of 157 patients was taken and features were extracted. The support vector machine classifier was applied

on integrated gene-expression and protein network dataset. Accuracy was calculated as a result of performance parameter and SVM performs well on this approach.

Cheng Wang [37] et al. proposed a prognosis model by integrating genomic and transcriptomic profiles for the detection of prognostic effect on breast cancer. Sample of 810 patients was taken TCGA dataset. Single-nucleotide polymorphism and genes related to breast cancer were extracted and divided into low survival and high survival patients. The experiment was performed and results showed that the proposed approach works well with the Area Under Curve (AUC) value of 0.79.

Siya Hen et al. [38] proposed a more practical strategy called Multiple Kernel Learning on omics data comprising miRNA expression, copy number variation, and DNA Methylation dataset. Sample of 10,000 patients was taken with 30 sub-types of cancer. Among 630 patients are selected having miRNA, copy number profiles and DNA methylation profiles and features were extracted. Once features got selected multiple kernel learning was applied. The results showed that MKL performed better as compared to Random Forest (RF) and Neural Network (NN).

A more practical strategy was proposed by Mohamed Amgad et al. [39] on genomic and histology data with Convolutional Neural Network (CNN) to predict cancer. The experiment was performed and results showed that the proposed approach performed best that other machine learning models.

Mo Li et al. [40] proposed another approach for breast cancer using a convolutional neural network with extreme learning machine algorithm. Sample of 400 patients was taken and the experiment was performed. The result showed that ELM performs best with a mean accuracy of 76%.

All the work done by various authors on genomic data, pathological dataset and integrated dataset is shown in table 2.1 below.

Table 2.1. Related Work

Author	Dataset	Methodology	Results	Year
Ashraf Abou Tabl et al. [19]	Gene Expression	SVM, Naïve Bayes, Random Forest	Accuracy	2019

Hongyue Dai et al. [20]	DNA Microarray	Random Forest	Accuracy, Sensitivity, Specificity, Area under the curve, MCC	2002
Lu Zou et al. [21],	Gene expression	PCA with autoencoder neural network	Concordance Index, AUC	2018
Zhi Huang et al. [22]	Gene expression	Machine Learning framework using neural network	Accuracy	2019
Chen Peng et al. [23]	Genomic	Deep Learning Based Capsule Network	Accuracy, Sensitivity, Specificity, Area under the curve	2019
Cheng Fan et al. [24]	Genomic	Logistic Regression, COX models	AUC, Concordance index	2015
Patrick Murigu Kamau Njage et al. [25]	DNA	SVM, Random Forest, Logic Boost	Area Under Curve	2019
Chaudhary et al. [26]	Genomic	SVM, RF, Linear Discriminant Analysis, Gradient boosting, regression tree	Area Under Curve	2019
J.S. Marron et al. [27]	Genomic	SVM, Random Forest	Accuracy	2003
Jan Ziang et al. [28]	Pathological Images	Convolutional Neural Network (CNN)	Accuracy	2016
Ce Zhang et al. [29]	Whole slide images	ML algorithms	AUC, Sensitivity, Specificity	2016

Jun Xu et al. [30]	Histopathology Images	Stacked Sparse Autoencoder	Precision, Recall, F-measure	2015
Ajay et al. [31]	Whole Slide Images	CNN	Positive Predictive Value, Negative predictive value	2017
Dayong Wang et al. [32]	Histopathology	Deep Learning Framework	Area Under Curve	2016
Olivier Gaveart et al. [33]	Clinical + Microarray	Bayesian Networks	Accuracy, AUC	2006
DongDong Sun et al. [34]	CNV + gene expr + clinical	Multimodal Deep Learning Neural Network	Sensitivity, Specificity, precision, recall, f-measure	2018
Hui Li et al. [35]	Genomic + Radiology	LASSO and Logistic Regression	Area Under Curve	2015
Hung-Wen Chui [36]	Gene Expr + Protein Expr	SVM	Accuracy	2017
Cheng Wang [37]	Genomic + Transcriptomic	Prognosis Model	AUC, Concordance Index	2019
Siya Hen et al. [38]	Gene Expr + CNV+ DNA	Multiple Kernel Learning	Area under curve	2019
Mohamed Amgad et al. [39]	Genomic + Histology	Convolutional Neural Network	Concordance Index	2018
Mo Li et al. [40]	Genomic	CNN with Extreme Learning Machine	Accuracy, Sensitivity, Specificity, AUC, Precision, Recall	2019

2.5 Summary

This part of the literature survey has clearly defined the causes and symptoms of breast cancer. The use of machine learning algorithms is clearly explained in the next section.

Further work done by various authors on simple genomic data, pathological data, and their integration is discussed. Based on this, we proposed a framework called GPELM that Genomic Pathological Extreme Learning Machine on integrated genomic and pathological images dataset for survival prediction of breast cancer patients.

Chapter 3
Research Gaps and Problem Statement

3.1 Research Gaps

This section gives a brief description of the gaps which were encountered during the review process of breast cancer survival prediction using machine learning.

i) In machine learning, feature extraction is an important part of any prediction. So, deep learning needs to be employed for feature extraction of multidimensional datasets [53].

ii) Handling missing value is one of the problems in multidimensional datasets. So, there is a need for more effective techniques to be developed to handle missing values in multidimensional datasets [54].

iii) There is a lack of deep learning algorithms for images dataset. Extracting features from images and integrating them with genomic dataset require the development of more effective models to improve the predictive performance of cancer patients [55].

iv) There is a need to develop ensemble models to effectively improve the performance prediction of survival of breast cancer patients [56].

v) The integrative analysis gets more complex in the case of the imbalanced dataset which develops a need of developing more effective techniques for handling them [57,58].

3.2 Problem Statement

Nowadays, breast cancer affected a lot of people because of various causes which cannot be determined easily. It is very difficult to treat cancer in the later stages of cancer, so early detection is become a necessity to save cancer patients. The test performed to detect cancer results in different number of values comprising thickness of the clump, size of the tumor, variability in the tumor and so on which is very difficult for the doctor to interpret the results. So, machine learning is the need for early detection and prediction of breast cancer. Earlier, predictive models used only clinical data, which gave moderate results due to the heterogeneity of different causes of breast cancer. Predicting cancer only on the basis of

clinical data and genomic data seems to be insufficient. Therefore, there is a need to consider a variety of data to predict more accurate results.

3.3 Objectives

- To study existing tools and techniques for breast cancer prediction.
- To select features from integrated genomic data and pathological images data for breast cancer.
- To develop breast cancer prediction model using the optimal feature subset.
- To test and validate the model using real-life data.

3.4 Methodology

The purpose of this research is to improve the survival prediction performance of breast cancer patients. This is achieved by using the proposed GPELM framework which is defined in the further sections. The methodology followed to predict survival of breast cancer patients is shown in Figure 3.1 and is described below in details:

Step 1: TCGA dataset including genomic dataset and pathological images dataset of 578 patients have been taken and downloaded using TCGA Bio-links package in R.

Step 2 (a): For the genomic dataset, data cleansing is performed which involves removing of duplicate and NA values from the dataset using k-mean clustering.

Step 2 (b): For pathological images, image tiling is done from which densest slides are selected using bftools.

Step 3: Feature Selection is performed for selecting the best informative features. For genomic, it is performed using FSelector and for pathological images, it is performed using cell profiler.

Step 4: Six machine learning models comprising ELMBJ, ELMBJEN, ELMCOX, ELMCOXEN, ELMCOXBOOST, ELMBOOST available in survelm library are selected for integration and training for breast cancer survival prediction.

Step 5: Performance is evaluated using various performance parameters comprising Sensitivity, specificity, accuracy, precision, hazard ratio, concordance index, AUC and AUPR.

Step 6: Finally, we get results in terms of patients having low survival or high survival and in terms of accuracy, sensitivity, precision, MCC, AUC, AUPR, hazard ratio, and concordance index.

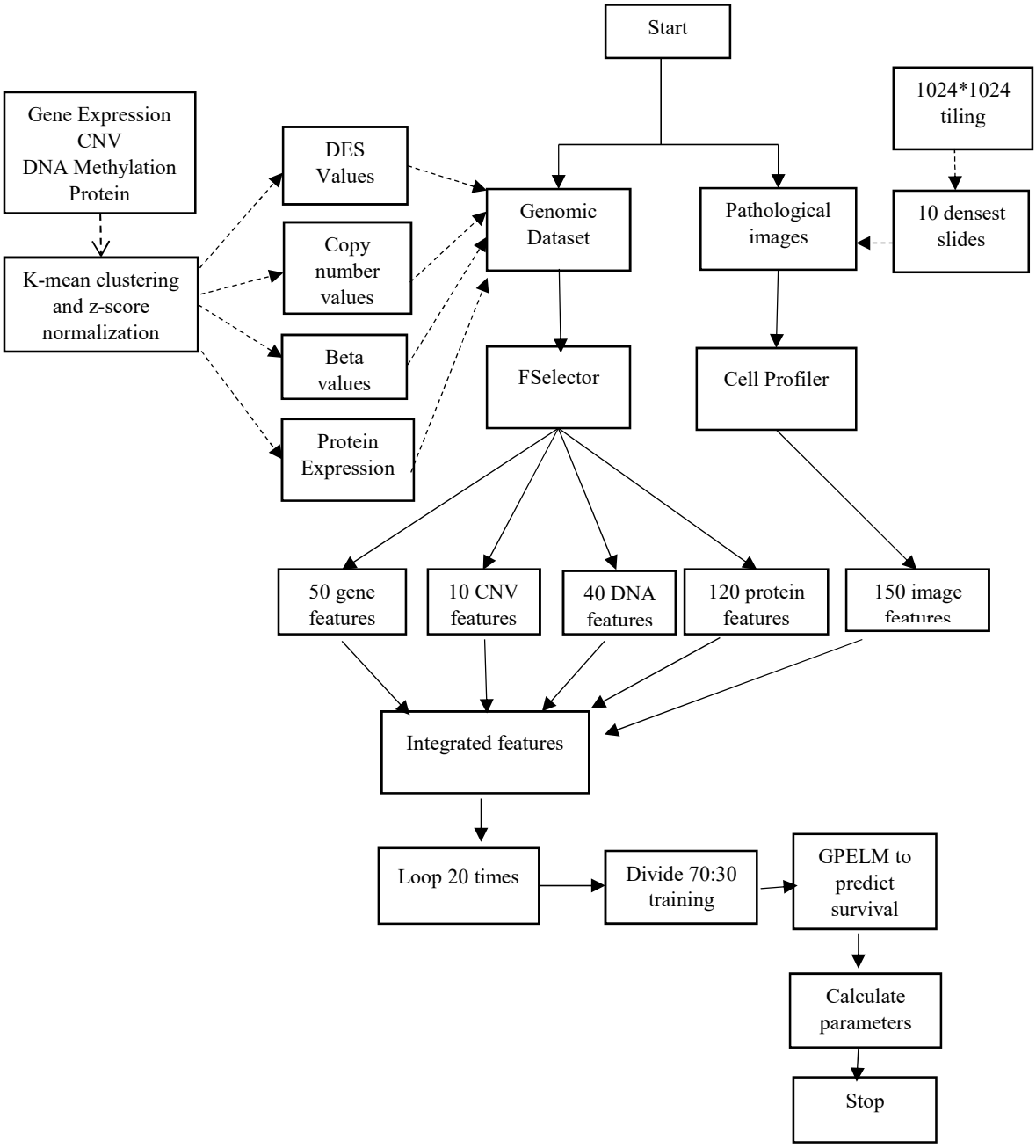

Figure 3.1. workflow Diagram-Methodology

3.5 Summary

This chapter concentrates on research gaps. Problem is formulated based on the research gaps. Objectives are set and the methodology to achieve the set objective is discussed.

Chapter 4
Proposed Framework

This chapter describes the proposed framework Genomic Pathological Extreme Learning Machine (GPELM). The framework is implemented on integrated dataset of genomic and pathological images for breast cancer survival prediction. A detailed description of the proposed approach is provided in this chapter. In the presented work, "Cancer Genome Atlas TCGA dataset" [41] is used.

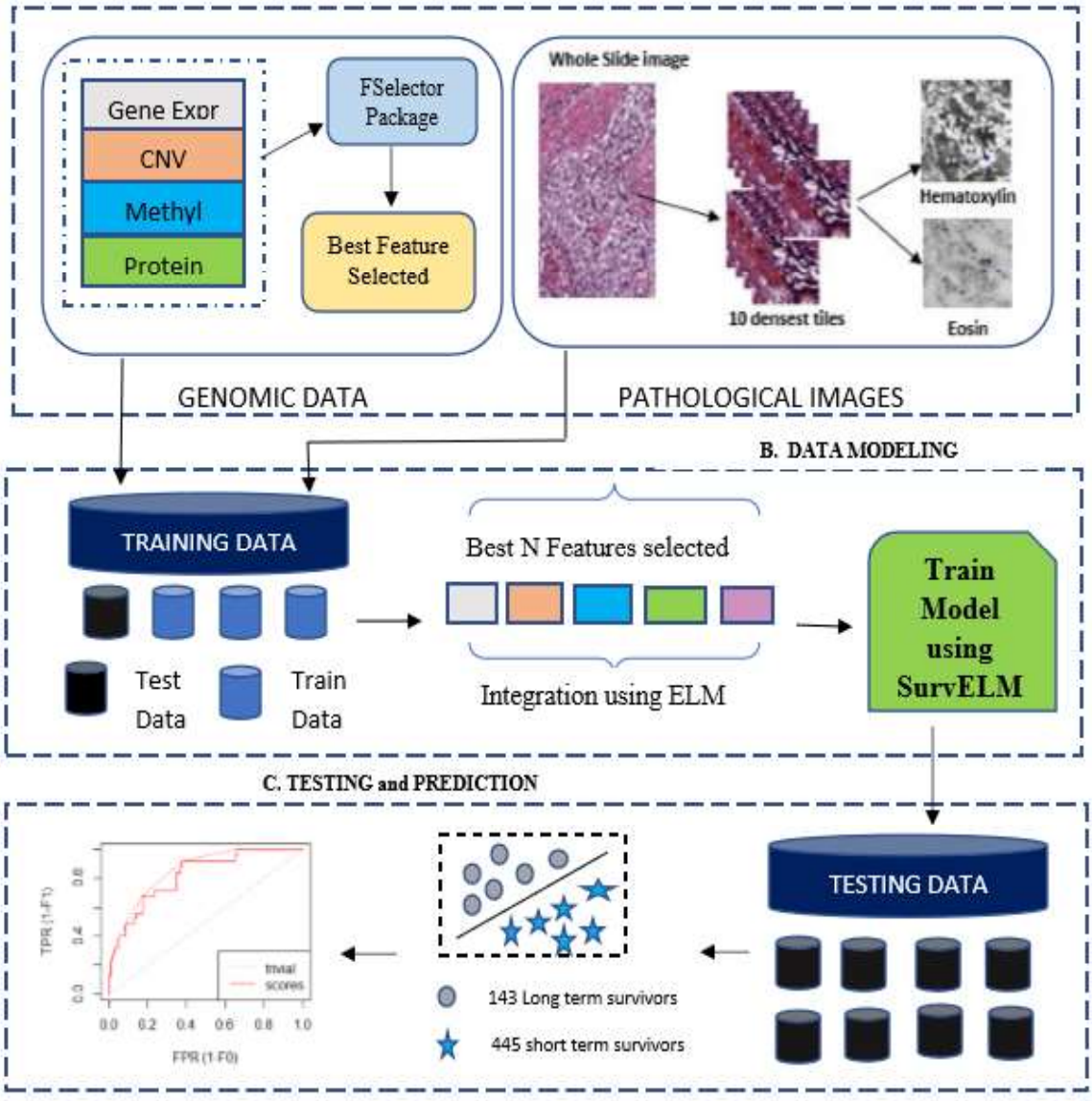

Figure 4.1. Overview of Proposed Work

Different stages are used to build the architecture which consists of data preparation, integration and use of machine learning models for early survival prediction of breast cancer patients as shown in figure 4.1 above.

The proposed framework GPELM, which is an invariant of the Extreme Learning Machine (ELM). This package comprises of ensembling of six different models. These models consist ELM with Buckley James estimator (ELMBJ), Ensemble of ELMBJ (ELMBJEN), ELM with regularized Cox model (ELMCOX), Ensemble of ELMCOX (ELMCOXEN), ELM with gradient-based boosting (ELMBOOST) and ELM with likelihood-based boosting (ELMCOXBOOST). The dataset has integrated genomic data (gene-expression, copy number alteration, DNA methylation, protein expression) and pathological images data for breast cancer survival prediction. These six models are present in the survELM package. At first, genomic data and pathological images dataset are taken which is preprocessed and the feature is extracted as described in Stage A of figure 4.1 above. Once the features are selected, the best features are integrated with the help of extreme learning machine defined in stage B. After that model training is done to train the model on survival data defined in stage B. In stage C, the model was tested on testing data and results are obtained which define 143 patients as a long-term survivor and 445 as short-term survivors.

4.1 Data Preparation

We collect the data from the Cancer Genome Atlas portal (TCGA). For this research, we take genomic data comprising gene expression, copy number alteration, DNA methylation, and protein expression data) and pathological images data. But each subtype contains a different number of patients. For example, for gene expression, there are 1250 samples and for images, there are 1010 samples. So, for all of them, we use Venn diagram to find the common patients and we get 578 valid female patients. The threshold value of 5 years is taken and the patients are classified into low survival having label value 0 and high survival having label value 1 the total no of patients having low survival is 445 and patients having low survival are 133. It is shown in table 4.1 below. The average age at which the diagnosis takes place is 57.80 and the average survival time for which the patient is going to survive is 40.5 months.

Table 4.1. Properties of Breast Cancer Dataset

Total Patients	578
Patients having Low-Survival	133
Patients having High-Survival	445

4.1.1 Preprocessing of Genomic Dataset

Genomic Data consists of gene expression, copy number alteration, DNA methylation, and protein expression dataset. To download the data, we simply use the TCGA Bioconductor package in the R source code [42]. Initially, we have the raw data which is very difficult to process. That is for gene expression we have 58000 transcriptome values, for copy number alteration we have 40000 copy number values with different chromosome which also includes duplicate values. For DNA methylation we have 70000 beta values and for protein expression, we have 20000 protein values. The size of the dataset is in terabytes. Handling this data is quite difficult. So, we need to preprocess it by using different techniques. Preprocessing involves data cleaning, data integration and normalization of data [43]. The preprocessing of genomic data is described below:

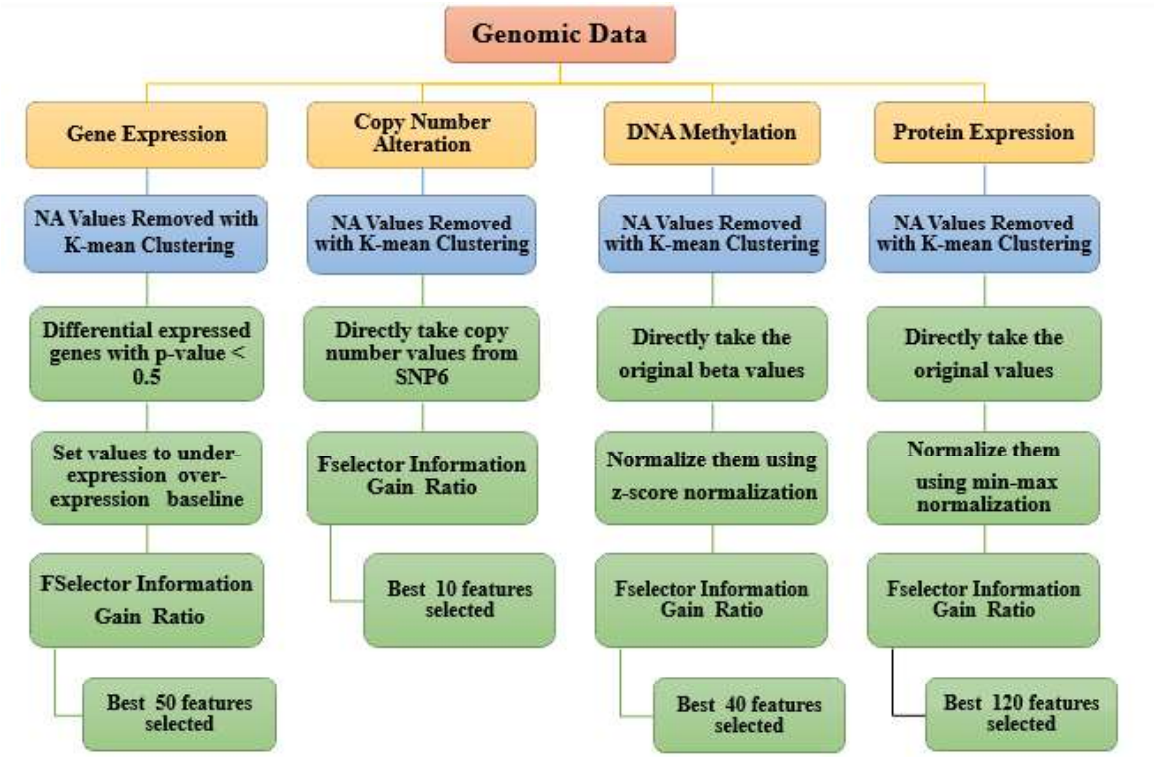

Figure 4.2. Preprocessing of Genomic Dataset

i. Gene Expression

Gene Expression data defines how the information is generated from the gene and how to use that information for making a useful gene product. The human body is made of a cell. Each cell contains miRNA data which is used to produce information. The flow of information starts with DNA. From there it moves to RNA and then to protein. this conversion of DNA to RNA is known as transcriptome data which is used in our research. In every second, thousands of transcripts are produced by each cell [16]. These transcripts are responsible for affecting the activity of the body. For gene expression, first 10% NA values are removed and then remaining empty values are removed using the k-means clustering algorithm. Then the differentially expressed genes are calculated using p-value having threshold value less than 0.5. After the values are divided into under-expression having value -1 if the value of gene is less than 0, overexpression having value 1 if the value of gene is positive and baseline having value 0 if the value of gene is 0 [44]. This is shown in figure 4.2 above.

ii. Copy Number Alteration

Copy Number Alteration data is a very important component of genome data and is described as a DNA segment of one or more kilobase as it is a variable number compared to gene-expression. This variation leads from several insertions, deletions, and update on the chromosome number which leads to a large variety of data [17]. These variations are implicated in somatic cells which leads to cancer. For copy number variation, first 10% NA values are removed and remaining empty values are removed with a k-mean clustering algorithm shown in figure 4.2 above. After that, directly the linear copy number values from Affymetrix SNP6 are selected and used for further process.

iii. DNA Methylation

In DNA Methylation, the methyl group is converted to cytosine ring of DNA at the 5^{th} position [18]. It modifies the function of genes and affects gene expression data. For DNA methylation,10% NA values are removed and remaining empty values are removed with a k-means clustering algorithm. We then directly take the original beta values from the remaining dataset and simply normalize them with z-Score normalization as shown in figure 4.2 above.

iv. Protein Expression

Similarly, Protein Expression data defines how the proteins are modified in the cells of a human body. Blueprints of proteins are stored in DNA and further decoded to produce RNA. RNA produces information in the form of proteins. For Protein Expression,10% NA values are removed and remaining empty values are removed with a k-means clustering algorithm. As shown in figure 4.2 above, we then directly take the original values from the remaining dataset and simply normalize them with min-max normalization.

4.1.2 Preprocessing of Pathological Images

Pathological images are whole slide images in which a glass slide is converted in a digital image which can be managed or analyzed on a computer screen. Tissue slide of the affected area is taken which can be seen in a microscopic environment in order to get the information. We can also generate numeric values of the area using various algorithms. For pathological images, Hematoxylin and Eosin whole slide image is used download from TCGA which are then tiled into 1024*1024 size image with the help of Bioconductor bftools [45]. When we tile the image then we get thousands of tiled images in gigabytes. Working with these large size images is quite difficult, so we select 10 tiles which are dense among these all as shown in figure 4.3 below.

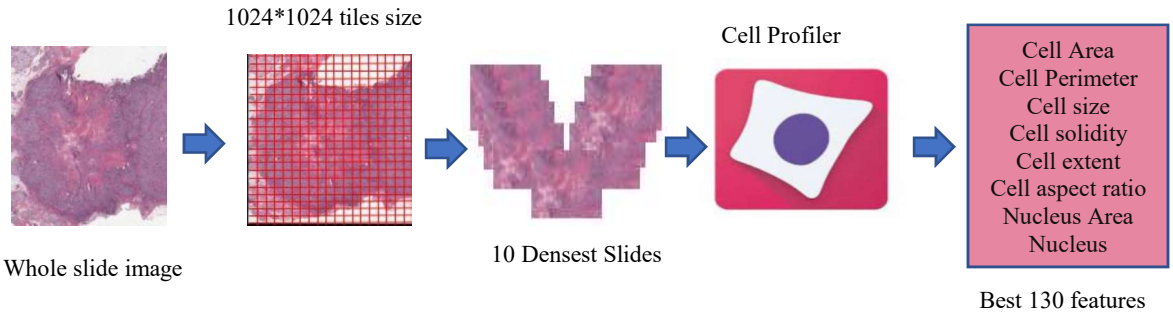

Figure 4.3. Preprocessing of pathological images

4.2 Feature Selection

Feature Selection is the process of selecting the most relevant features which are of great importance for our prediction process [46]. One of the most effective feature selection technique used is to select redundant and irrelevant features and then remove them [47, 48]. Feature Selection is divided into filter method which ranks the most important features

and contains information gain, gain ratio techniques [49] shown in figure 4.4 below, wrapper method which takes a subset of features, then train them, check performance and again take another subset and so on [50], Embedded method which are same as wrapper method but they use learning model for selection purpose [51]. The proposed framework uses filter method information gain ratio approach for selecting the best features for genomic data and cell profiler for pathological images dataset.

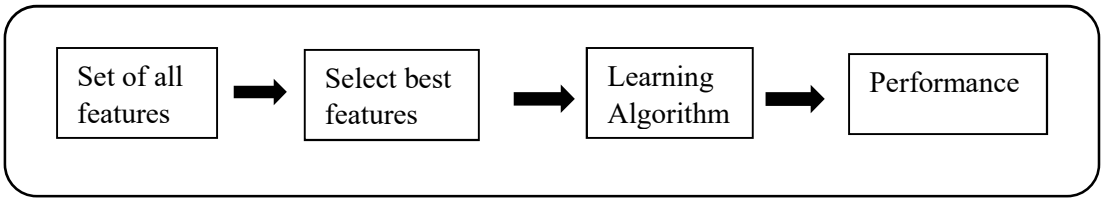

Figure 4.4. Filter Method

After preprocessing of genomic data, we are left with 15000, 25000, 16000, 215 features for gene expression, copy number, DNA methyl, and protein expression dataset. Working with such huge features is quite difficult. For the selection of most relevant features, we make use of FSelector package which selects the most informative features using information gain ratio to avoid overfitting [52,53]. FSelector package follows the following algorithm for selecting the most informative genes.

Input: Preprocessed Data

Output: Most informative features

1 **For each** feature feature_i
2 Set **Weight_i =** information.gain (feature_i)
3 Add **Weight_i** to weight list
4 Sort Weight list
5 Set **subset** = cutoff.k (weights, k)
6 Select subset as features

end

Algorithm 4.1. FSelector Package Algorithm

This algorithm selects top 50 features for gene expression, 10 most informative copy number values, 40 features for DNA Methylation and 120 best features for protein expression data. The features are selected on the basis of their importance in the proposed approach. CNA impact in our dataset is very less so we take only 10 features from the

whole values. Similarly, the impact of protein expression dataset is very high for survival prediction. So, more features are selected for them as compared to other subtypes. This value range is selected on the basis of weights value. As genomic data is passed to FSelector package, then it ranks features highest to lowest and select the top-ranked features based on the cut-off value [53]. For Pathological images, we are left with 1991 image features. In order to extract important image features Cell Profiler is used. After profiling, we get 130 images features including cell nuclei, cell nuclei, cell shape, cell size, no of tiles, no of cells in images tiles, the texture of image tiles and so on shown in figure 6 above. The total no of features extracted from the dataset is described in table 4.2 below:

Table 4.2. No of features of Breast Cancer Dataset

Sub Type	Features Before	Feature after extraction
Gene Expression	16000	50
Copy Number Alteration	25000	10
DNA Methylation	16000	40
Protein Expression	215	120

4.3 Machine Learning Models for Training

The proposed approach GPELM consists of the survELM package for integrating genomic and pathological images data and for predicting the survival outcome. The extracted features of both genomic data and pathological images are taken, in which gene expression, DNA methylation, Copy Number Alteration, and Protein Expression takes miRNA (50 values), Beta values (40 values), Linear Copy numbers (10 values) and protein data (120 values) as features and pathological images includes cell nuclei, cell shape, cell size, color red blue green (RBG) and so on up to 196 features. Due to this high dimensionality of data, Kernel Extreme Learning Machine (GPELM) has become a powerful choice for more efficient results.

4.3.1. Extreme Learning Machine

Extreme Learning Machine works on single-hidden layer feed forward neural network in which input weight value is chosen randomly and output is generated accordingly. From past years, ELM has been used performed on high dimensional datasets for survival analysis [29]. It works as follows:

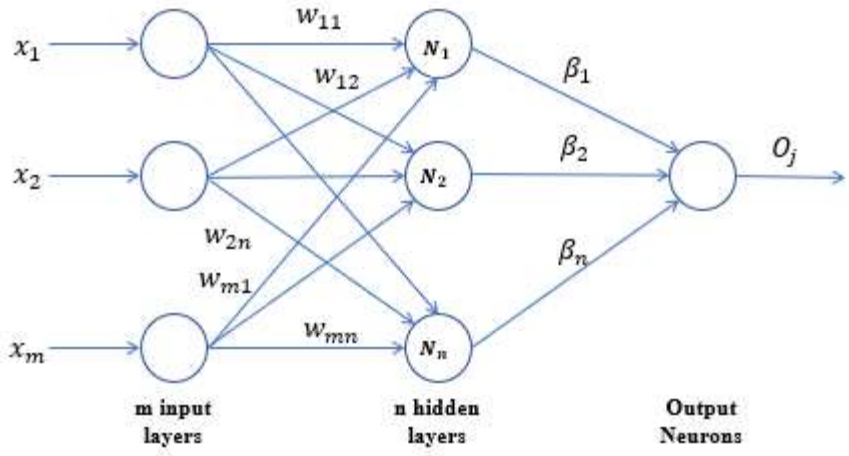

Figure 4.5. Structure of ELM

For a given training sample, $\{x_i, y_i | \ x_i \in R^P, y_i \in R^m\}_{i=n}^n$, if n defines total observations, p gives the dimension of covariates, y_i defines the target for each observation, then ELM with n hidden layers can be written as:

$$f_L(x) = \sum_{i=1}^{L} g\left(x, w_i, b_i\right)\beta_i = h(x)\beta \qquad (4.1)$$

Here, g defines activation function, w_i defines input weights, b_i defines the bias variable, $h(x)$ defines hidden layers, β defines the output target variable. The hidden layer for ELM can be expressed as:

$$\mathrm{H} = \begin{bmatrix} h(x1) \\ h(x2) \\ . \\ . \\ h(xn) \end{bmatrix} = \begin{bmatrix} g(w_1, b_1, x_1) \ldots g(w_L, b_L, x_1) \\ g(w_1, b_1, x_2) \ldots g(w_L, b_L, x_2) \\ . \\ . \\ g(w_1, b_1, x_n) \ldots g(w_L, b_L, x_n) \end{bmatrix}_{(n*l)} \qquad (4.2)$$

And the target matrix is given by:

$$\mathrm{Y} = \begin{bmatrix} y_1^T \\ y_2^T \\ . \\ . \\ y_n^T \end{bmatrix} = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ & . & \\ & . & \\ y_{n1} & \ldots & y_{nm} \end{bmatrix} \qquad (4.3)$$

The output weights can be solved with the given equation:

$$\beta = H^Y \left(\frac{I}{C} HH^Y\right)^{-1} Y \qquad (4.4)$$

Where I is a n*m matrix.

Kernel ELM can be defined by the following equation:

$$K(x_i, x_j) = \mathrm{h}(x_i).h(x_j) \quad (4.5)$$

Kernel ELM with L supporting vectors can be given by:

$$fL(x_i) = \sum L_j = 1\, K(x_i, x_j)\beta_j, \;\; i = 1,2,\dots,n \quad (4.6)$$

Where $fL(x) = K_{n\times l}\beta$

So, based on the above criteria, the extracted features from data preprocessing stage and feature selection stage, are integrated with the help of kernel Extreme Learning Machine. The kernel used in the whole process is of the linear type which is given by the following equation:

$$k(x, y) = \sum_i^N \propto_i y_i\, (x_i^T x) + b \quad (4.7)$$

Here, (x, y) belongs to training data, b is a constant with adjustable parameter $\propto$.

Once the features are integrated, we train them using survELM package which consists of six models, Extreme Machine Learning Algorithms (ELM) with Buckley James estimator and its ensemble, ELM with cox regularized models and its ensemble and with likelihood and gradient boosting.

4.3.2 Extreme Learning Machine with Buckley James Estimator

Buckley-James estimator was introduced in 1979 by Buckley and James which make use of least square estimation method for censored data. Buckley-James estimator is explained as

Suppose a random variable Y forms a linear regression on some covariate X, then

$$Y = x\beta + \epsilon, \quad (4.8)$$

Where x is a vector of 1×p with a constant β and having e as a zero mean and finite variance random variable [30], Y represents the survival time or an event.

The survival time is estimated with Buckley-James estimator and then ELM is applied for survival analysis of patients. This whole process is named as ELMBJ. For more perfect results, an ensemble of extreme learning machine with Buckley-James Estimator is

performed which is known as survival ensemble of extreme learning machine with Buckley-James estimator (ELMBJEN).

4.3.3 Regularized Cox with Extreme Learning Machine

In this, the linear version of the cox model is replaced with a non-linear elm neural network and as a result, the coefficient will be obtained [29]. Assume x as a variable set having a total number of p covariates on a dataset D with training samples n$\times$ $(p + 2)$, where D belongs to $(\tau_{,}\delta_i,\ x_i)$, i=1,2….,n. In case of right censored data, $\tau_i = \min(T_i, C_i)$, here T_i is true survival time, C_i is censored status and δ_i is censoring indicator [28]. The hazard of a patient can be written as:

$$(\lambda_i(t) = \lambda_0(t)\exp(f(x_i)) \qquad (4.9)$$

Here $f(x_i)$ is a function of covariate x_i, and for traditional cox model, $f(x_i) = x_i\beta$. Also, a random-forest based ensemble is provided called ELMCoxen for its stable result.

4.3.4 Extreme Learning Cox model with Gradient Based Boosting

In this, we apply the elm model in boosting the environment to get the survival data. In this, two types of boosting named gradient boosting and likelihood-based boosting are used. All these are present in a package called survELM package [30].

4.3.5 Ensemble Modeling

The working for the whole is explained here in this section. The extracted features from the datasets using feature selection methods are integrated with the help of kernel Extreme Learning Machine. The kernel used in the whole process is of the linear type which is given by the following equation:

$$k(x, y) = \sum_i^N \propto_i y_i (x_i^T x) + b \qquad (4.10)$$

Here, (x, y) belongs to training data, b is a constant with adjustable parameter $\propto$.

For training, first, we applied the Buckley James estimator with Extreme Learning Machine (ELMBJ) in which ELM integrates the data by extracting the important features from all the datasets and then Buckley James estimator will predict the survival outcome. In this, the kernel used is of linear type and value of alpha used is 0.5. In order to improve the performance of this ELMBJ, an ensemble version of this model that is Ensemble of

Extreme Learning Machine with Buckley James Estimator (ELMBJEN) is applied and obtained the results. This ensemble method uses 100 ELM base survival models for integrating the data and predicting the performance. After, Regularized Cox with Extreme Learning Machine (COXELM) is method is used for integration by taking linear kernel with an alpha value of 0.5. This method is enhanced with its ensemble version that is Ensemble of Regularized Cox with Extreme Learning Machine (ELMCOXEN). In this, it uses a linear kernel by taking an alpha value of 0 with 100 base models. After that, we make use of Extreme Learning Machine with Gradient Boosting (ELMBOOST) and Likelihood-based Boosting (ELMCOXBOOST) with an alpha value of 0.5 by taking a linear kernel. For more efficient results, the whole process is iterated 20 times and the result is compared with different baseline models given by various authors including SuperPC, RSF, Survival Regression (Survreg) and BoostCI. This whole is implemented in R studio. The results show that the ELMBJ, ELMCOX produces sufficient results and their ensemble version (ELMBJEN, ELMCOXEN), ELMBOOST and ELMCOXBOOST slightly increase the performance by a factor of 0.2. In order to see the performance, different parameters are used which we get after comparing them with different models. The area under the curve for each is achieved for each model. The performance parameters calculated are as follows:

i. $$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (4.11)$$

ii. $$\text{Specificity} = \frac{TN}{TN+FP} \quad (4.12)$$

iii. $$\text{Precision} = \frac{TP}{TP+FP} \quad (4.13)$$

iv. $$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.14)$$

v. $$\text{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TN+FP)\times(TN+FN)\times(TP+FP)}} \quad (4.15)$$

Here TP stands for True Positive, FP for False Positive, TN for True Negative and FN for False Negative which can be obtained by drawing a confusion matrix. It is used to define the performance of the model by taking actual values and predicted values and is displayed in the form of a matrix as shown below in table 4.3:

Table 4.3. Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

vi. CI $= \frac{1}{n} \sum_{i\epsilon\{1..n|\delta_i=1\}} \sum_{s_{j>}s_i} I\left[X_i\hat{\beta} > X_j\hat{\beta}\right]$ (4.16)

Here n is total no of comparison, I represent the indicator and s represents the actual result.

4.4 Summary

In this chapter, proposed framework GPELM using six machine learning models comprising ELM with Buckley James estimator (ELMBJ), Ensemble of ELMBJ (ELMBJEN), ELM with regularized Cox model (ELMCOX), Ensemble of ELMCOX (ELMCOXEN), ELM with gradient-based boosting (ELMBOOST) and ELM with likelihood-based boosting (ELMCOXBOOST) is explained in detail. The performance parameters required are also discussed.

Chapter 5

Implementation and Experimental Results

This chapter discusses the experimental setup for the GPELM framework along with the implementation and results.

5.1 Experimental setup

5.1.1 Minimum Software and Hardware requirements

Table 5.1 H/W and S/W Requirements

1.	Processor	32 bit
2.	RAM	4 GB
3.	Hard Disk	80 GB
4.	Operating System	Windows 7
5.	Programming Language	R (Rattle)
6.	Platform	R Studio

5.1.2 Extreme Learning Machine with Buckley James Implementation

We implemented machine learning algorithm with the help of the R language. For ELMBJ implementation we install survelm package and also other libraries including hmeasure for confusion matrix, survival for checking the survival of patients, plotRoc for plotting the curve. After that data read and write operations are performed on the dataset and the performance of ELMBJ is determined in terms of various performance parameters. The ELMBJ can be implemented in R with the following functions including libraries, package, and method.

5.1.3 Ensemble of Extreme Learning Machine with Buckley James Implementation

For ELMBJEN implementation, we install survelm package and also other libraries including hmeasure, survival, plotRoc for plotting the curve. After that data read and write operations are performed on the dataset and the performance of ELMBJEN is determined in terms of various performance parameters. The ELMBJEN implementation in R with the various functions including libraries, package, and method is shown below.

5.1.4 Extreme Learning Machine with Regularized Cox

For ELMCOX implementation, we install survelm package and also other libraries including hmeasure, survival, plotRoc for plotting the curve. After that data read and write operations are performed on the dataset and the performance of ELMCOX is determined in terms of various performance parameters.

5.1.5 Ensemble of Extreme Learning Machine with Regularized Cox

For ELMCOXEN implementation, we install survelm package and also other libraries including hmeasure, survival, plotRoc for plotting the curve. After that data read and write operations are performed on the dataset and the performance of ELMCOXEN is determined in terms of various performance parameters.

5.1.6 Extreme Learning Machine with Gradient Boosting

For ELMBOOST implementation, we install survelm package and also other libraries including hmeasure, survival, plotRoc for plotting the curve. After that data read and write operations are performed on the dataset and the performance of ELMBOOST is determined in terms of various performance parameters.

5.1.7 Extreme Learning Machine with Likelihood Boosting

For ELMCOXBOOST implementation, we install survelm package and also other libraries including hmeasure, survival, plotRoc for plotting the curve. After that data read and write operations are performed on the dataset and the performance of ELMCOXBOOST is determined in terms of various performance parameters.

5.2 Results

In order to predict the survivability of Breast cancer patients and to predict the outcome of various machine learning models, we use TCGA dataset. To download the data, we simply use the TCGA Bioconductor package in the R source code. Initially, we have dataset in Raw form. That is for gene expression we have 58000 transcriptome values, for copy number alteration we have 40000 copy number values with different chromosome which also includes duplicate values. For DNA methylation we have 70000 beta values and for protein expression, we have 20000 protein values. The size of the dataset is in terabytes. Handling this data is quite difficult. So, we need to preprocess it by using different

techniques. The preprocessing of data is explained in section 4.1. After preprocessing we still left with 16000, 25000, 16000, 215 features of gene expression, Copy Number Alteration, DNA Methylation, and protein expression. In order to reduce them, we make use of FSelector Information gain ratio package which uses cutoff value to select the most informative genes from genomic dataset. After going through feature selection, we left with 50, 10, 40, and 120 features of gene expression, CNV, DNA Methylation, and protein expression.

Table 5.2. Summary of Breast Cancer Dataset

TCGA	
Characteristics	**Summary**
Total Breast Cancer patients	585
Gender	Male = 7, female = 578
Selected Patients	578
Short term survivor	445
Long term survivor	133
Average Age of Diagnosis	30-78 (approx. avg= 58)
Data selected	
Genomic Data	**Features selected with FSelector Package**
Gene Expression	50 features
Protein Expression	120 features
DNA Methylation	40 features
Copy Number Alteration	10 features
Pathological images	196 features through cell profiler

Similarly, for pathological image whole slide diagnose image is taken in which one single is up to 800MB. The total images size for all patients goes in terabytes. So, for that, we need to tile the images. When we tile a single image, we got hundreds of tiles. If we go for a single tile, numerous features can be obtained which results in large data for all tiles. So, we select only the densest images which are of our interest. After this whole process, we still left with 1991 features. So, for finding the important features cell profiler is used which reduces the features to 196 consisting of cell size, cell shape, cell perimeter, cell area, nucleus area, and so on. The total no of features extracted from the dataset is described in table 5.2 above.

For results, we applied all the six models given in survELM package to effectively predict breast cancer. We train the six models on high dimensional data by integrating genomic and pathological images and repeat the whole process 20 times. Results are displayed in table 5.3.

Table 5.3. Results of the proposed framework

Paper	Models	Sensitivity (95%)	Precision	Accuracy	AUC	AUPR	MCC	Hazard Ratio	CI
D Sun et al. [33]	SurvReg	0.17	0.49	0.74	0.69	0.55	0.41	0.98	0.59
	RSF	0.19	0.4	0.73	0.71	0.50	0.43	0.54	0.57
	SuperPC	0.20	0.46	0.76	0.72	0.42	0.42	0.65	0.53
	BoostCI	0.21	0.6	0.79	0.725	0.54	0.45	0.76	0.55
	GPMKL	0.28	0.62	0.80	0.80	0.68	0.50	0.44	0.60
Our Approach	**ELMBJ**	**0.30**	**0.63**	**0.83**	**0.83**	**0.75**	**0.52**	**0.14**	**0.63**
	ELMBJEN	**0.312**	**0.635**	**0.835**	**0.834**	**0.73**	**0.54**	**0.09**	**0.64**
	ELMCOX	**0.301**	**0.63**	**0.846**	**0.84**	**0.71**	**0.55**	**0.32**	**0.63**
	ELMCOXEN	**0.31**	**0.64**	**0.84**	**0.845**	**0.72**	**0.54**	**0.07**	**0.635**
	ELMCOXBOOST	**0.315**	**0.63**	**0.84**	**0.84**	**0.71**	**0.53**	**0.43**	**0.63**
	ELMBOOST	**0.32**	**0.64**	**0.85**	**0.85**	**0.75**	**0.56**	**0.05**	**0.64**

The area under curve value and other performance parameters are calculated for each model. The roc curves for all the six values are shown below. The dark solid line in the curve shows the actual values. It determines the curve between the true positive rate and false positive rate. It determines that how many patients having cancer is actually have cancer. ELMBJ gives an AUC value of 0.83 and its ensemble slightly increase the value by 0.5 that is its values is 0.835 and is defined in figure 5.1.

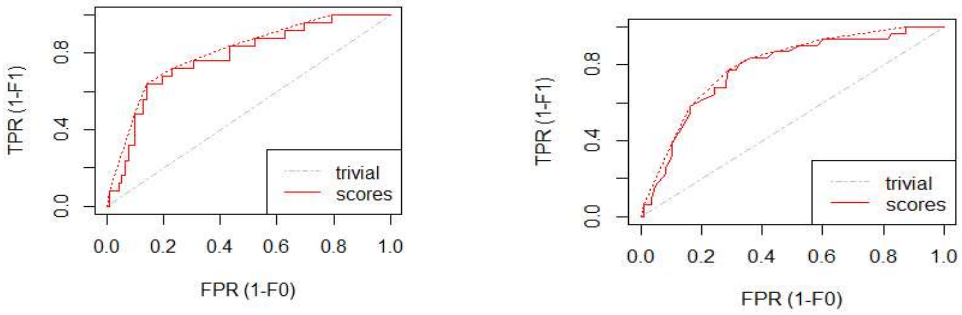

Figure 5.1. ROC Curve value for ELMBJ and ELMNJEN

Similarly, ELMCOX produces an AUC value of 0.84 and its ensemble give the value of 0.85. This is shown in figure 5.2 below.

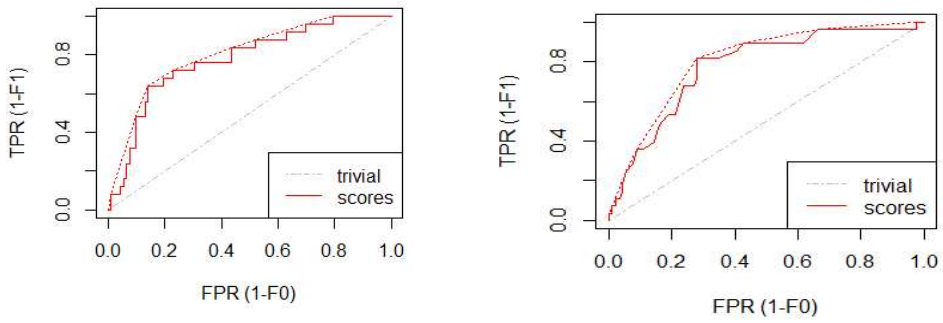

Figure 5.2. ROC Curve for ELMCOX and ELMCOXEN

For ELMCOXBOOST and ELMBOOST gives the value of 0.84 and 0.85 and is defined in figure 5.3. From results, we proved that from among all the six models, ELMBoost performs best with an AUC value of 0.85.

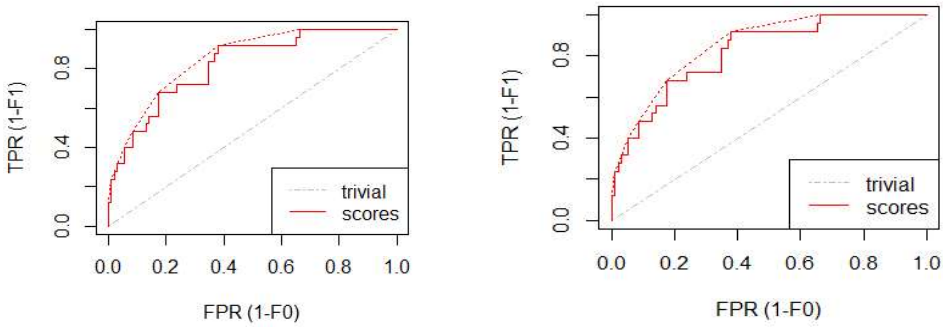

Figure 5.3. ROC Curve for ELMCOXBOOST and ELMBOOST

Along with GPELM, we also compare our results with the other survival models comprising cox models which includes survreg, RSF, SuperPC, BoostCI and, GPMKL. The results are shown in table 5.3 above. We also compare the results with the lung cancer results implemented using survelm library. The lung cancer data is simple clinical data. We implement this on integrated data and prove that our proposed framework will produce the best results with the best accuracy. Our proposed approach produces an accuracy of 83% and 83.5% for ELMBJ and ELMBJEN. For ELMCOX and ELMCOXEN produces an accuracy of 84% and ELMBOOST and ELMCOXBOOST produces an accuracy of 85% and 84%. Along with that, other parameters including sensitivity, precision, Matthews correlation coefficient, and concordance index is also calculated. Compared to the other

models, the proposed framework increased the result by 4%, 2%, 5%, 5%, 7%, 6% and 4% for sensitivity, precision, accuracy, AUC, AUPR, MCC and concordance index.

We also plot bar graphs based on precision, accuracy, sensitivity, and concordance index value for different models and our models to show the difference in results graphically. It is clearly visible the difference between different models. The values for BoostCI, Survreg, Random Forest and SuperPC are far less than the proposed framework and for GPMKL there is a slight difference by 5%.

The comparative performance for base paper models along with our proposed approach in terms of accuracy is shown in figure 5.4 below.

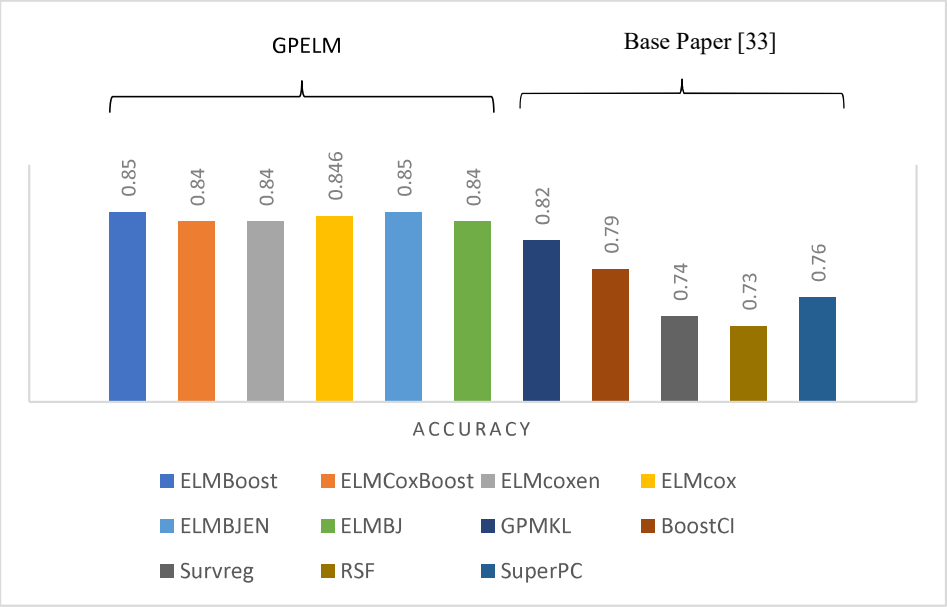

Figure 5.4. Bar graph for Accuracy.

The comparative performance for base paper models along with our proposed approach in terms of sensitivity is shown in figure 5.5 below.

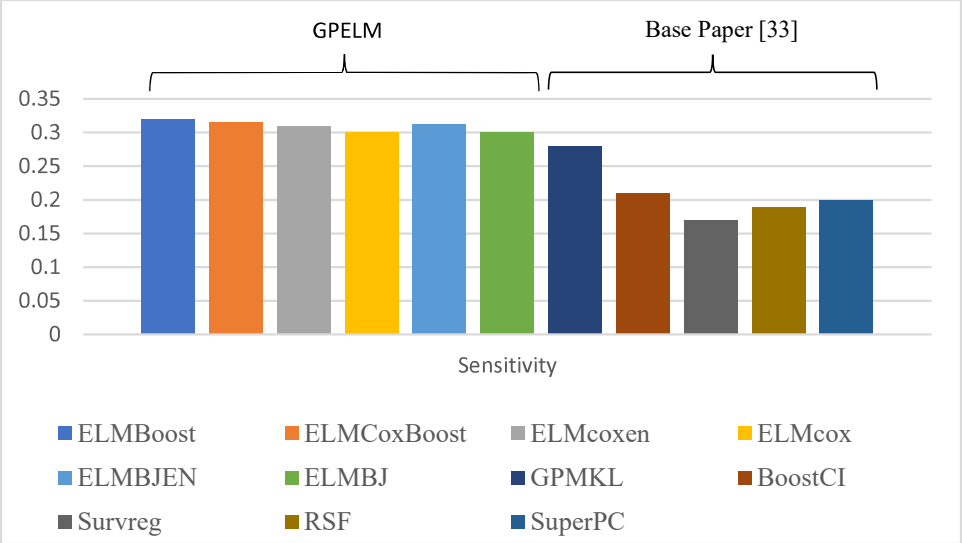

Figure 5.5. Bar graph for Sensitivity.

The comparative performance for base paper models along with our proposed approach in terms of precision is shown in figure 5.6 below.

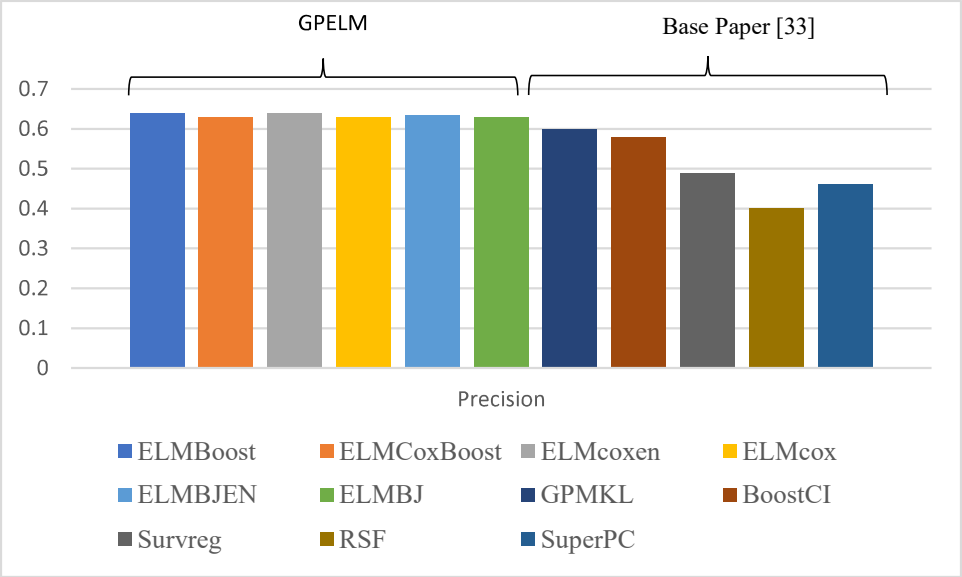

Figure 5.6. Bar graph for Precision.

The comparative performance for base paper models along with our proposed approach in terms of precision is shown in figure 5.7 below.

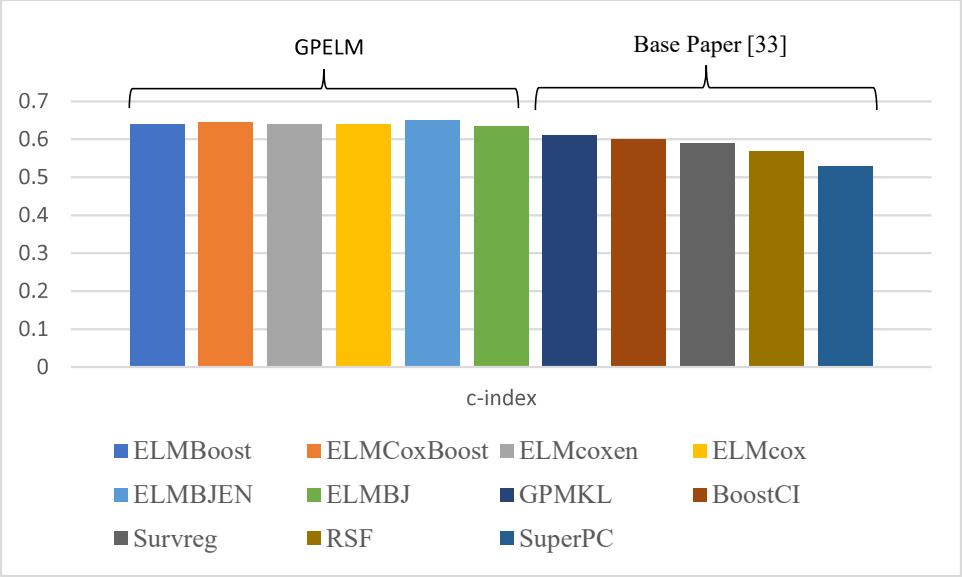

Figure 5.7. Bar graphs for CI

An individual experiment is implemented to discuss which part plays an important role in improving the results. To do this, every time one of gene expression, CNA, methylation, and protein is removed, and then we draw the ROC curve for comparison with the complete GPELM. We find that all the information is important for breast cancer survival prediction in which the highest role is played by gene expression and protein expression. Also, we implement the whole experiment on pathological images alone and then integrate it with one of the types of genomic data. The results obtained show that integration plays an important role. The reason why GPELM achieves the best performance is that genomic data and pathological images can provide predictive powers and could have a complementary relationship, and extreme learning machine is very efficient in predicting breast cancer survival time.

5.3 Summary

This chapter provides an implementation of machine learning algorithms and also discussed various parameters like accuracy, sensitivity, precision, AUC, AUPR, MCC, hazard ratio and concordance index. From results it is cleared that there is an average increase of 5% in performance for breast cancer survival prediction with 4%, 2%, 5%, 5%, 7%, 6%, and 4% improvement in each parameter comprising sensitivity, precision, accuracy, AUC, AUPR, MCC and concordance index.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

As breast cancer is the leading cause of death in females, so the focus is to refine the results of survival time prediction for breast cancer patients. In the present research, GPELM is proposed which includes six models comprising Extreme machine learning with Buckley James estimator and its ensemble, Extreme Learning Machine with regularized Cox and its ensemble, Extreme Learning Machine with gradient boosting and likelihood-based boosting. GPELM is applied to the integration of genomic and pathological images data in order to predict the survival of breast cancer patients. GPELM is experimentally analyzed using various parameters comprising, sensitivity, specificity, precision, accuracy, area under curve, area under precision-recall, Mathews correlation coefficient, hazard ratio, and concordance index value. After training of the model, the results are achieved for each model of survELM library. The results show that proposed approach works well with 85% accuracy using ELMBOOST, 84% using ELMCOXBOOST and ELMCOX, and 83% with ELMBJ, ELMBJEN, and ELMCOX by accurately predicting the survival of breast cancer patients. ELMBOOST performs best among all the six models. The results are also compared with the results of the base paper [33]. It is observed that the proposed approach shows improvement and increase in each performance parameter comprising sensitivity, precision, accuracy, AUC, AUPR, MCC and concordance index by 4%, 2%, 5%, 5%, 7%, 6%, and 4% respectively in breast cancer survival prediction. One more dataset for lung cancer has been tested using this package which is giving an equal performance. So, this package can also be applied to predict another type of cancers using clinical data, genomic data, images data as well as the integration of these data.

6.2 Future Scope

In the future, deep learning algorithms can be applied for further improvement and the work can be extended on bio-marker prediction and other datasets. Another research direction is to construct a multi-task learning system aiming to cancer susceptibility, cancer recurrence, and cancer treatment.

References

[1] L.M. Franks, and M. T. Natalie, eds. *Introduction to the cellular and molecular biology of cancer*. Oxford University Press, USA, 1997.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: a cancer journal for clinicians*, vol. 68, no.6, pp.394-424, 2018.

[3] "What is Breast Cancer", www.breastcancer.org. [Online] Available: https://www.breastcancer.org/symptoms/understand_bc/what_is_bc. [Accessed: 26 April, 2019].

[4] "Types of Breast Cancer", www.breastcancer.org. [Online] Available: https://www.breastcancer.org/symptoms/types. [Accessed: 26 April, 2019].

[5] "Diagnosis of Breast Cancer", www.cancer.ca. [Online] Available: http://www.cancer.ca/en/cancer-information/cancer-type/breast/diagnosis/?region=on. [Accessed: 30 April, 2019].

[6] E. Alpaydin, "Introduction to machine learning", MIT Press, 2009.

[7] G.B. Huang, Q.Y. Zhu and C.K. Siew, "Extreme learning machine: theory and applications", *Neurocomputing*, vol. 70, no. 1-3, pp.489-501, 2006.

[8] S. Ding, H. Zhao, Y. Zhang, X. Xu and R. Nie, "Extreme learning machine: algorithm, theory and applications" *Artificial Intelligence Review*, vol. *44,* no.1, pp.103-115, 2015.

[9] U. Potter, "A multivariate Buckley-James estimator", [Online] Available: https://www.demogr.mpg.de/papers/workshops/010830_paper01.pdf. [Accessed: 1 May, 2019].

[10] P. Grover, "Gradient Boosting from Scratch". [Online] Available: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d. [Accessed: 5 May, 2019].

[11] G. Huang, S. Song, J.N. Gupta and C. Wu, "Semi-supervised and unsupervised extreme learning machines", *IEEE transactions on cybernetics*, vol.44, no.12, pp.2405-2417, 2014.

[12] Alpaydin and Ethem, "Introduction to machine learning", *MIT press, Cambridge*, 2009.

[13] S.B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques", *Emerging artificial intelligence applications in computer engineering*, vol.160, pp.3-24, 2007.

[14] C.E. Rasmussen, "Gaussian processes in machine learning", *In Summer School on Machine Learning,* pp. 63-71, Springer, Berlin, Heidelberg, 2003.

[15] X. Zhu and A.B. Goldberg, "Introduction to semi-supervised learning", *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no.1, pp.1-130, 2009.

[16] R. Shyamsundar, Y. Kim, J. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, M. Rijn, D. Botstein, P. Brown and J. Pollack, "A DNA microarray survey of gene expression in normal human tissues*", Journal of Genome Biology,* vol.6, no.3, p.22, 2005.

[17] R. Redon, S. Ishikawa and K. Fitch, "Global variation in copy number in the human genome"*, Journal of Science,* pp.444-454, 2006.

[18] B. Jin, Y. Li and K. Robertson, "DNA Methylation Superior or Subordinate in the Epigenetic Hierarchy?", vol.2, no.6, pp.607-6017, 2011.

[19] A.A. Tabl, A. Alkhateeb, W. ElMaraghy, L. Rueda and A. Ngom, "A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer", *Frontiers in Genetics*, vol. 10, 2019.

[20] L.J. Van't Veer, H. Dai, M.J. Van De Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. Van Der Kooy, M.J. Marton, A.T. Witteveen and G.J. Schreiber, "Gene expression profiling predicts clinical outcome of breast cancer", *nature*, vol. 415, no. 6871, pp.530-536, 2002.

[21] D. Zhang, L. Zou, X. Zhou and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer", *IEEE Access*, vol. 6, pp.28936-28944, 2018.

[22] Z. Huang, X. Zhan, S. Xiang, T.S. Johnson, B. Helm, C.Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han and K. Huang, "SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer", *Frontiers in genetics*, vol.10, p.166, 2019.

[23] C. Peng, Y. Zheng and D.S. Huang, "Capsule Network based Modeling of Multi-omics Data for Discovery of Breast Cancer-related Genes", *IEEE/ACM transactions on computational biology and bioinformatics,* 2019.

[24] A. Prat, C. Fan, A. Fernández, K.A. Hoadley, R. Martinello, M. Vidal, M. Viladot, E. Pineda, A. Arance, M. Muñoz and L. Paré, "Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy", *BMC medicine*, vol.13 no.1, p.303, 2015.

[25] P.M.K. Njage, P. Leekitcharoenphon and T. Hald, "Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in shigatoxigenic Escherichia coli" *International journal of food microbiology,* vol. 292, pp.72—82, 2019.

[26] F.M. Alakwaa, K. Chaudhary and L.X. Garmire, "Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data", *Journal of proteome research,* vol.17, pp.337—347, 2017.

[27] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J.S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler and J. Demeter, "Repeated observation of breast tumor subtypes in independent gene expression data sets", *Proceedings of the National Academy of Sciences of the United States of America*, vol.100, no.14, pp.8418-8423, 2003.

[28] F.A. Spanhol, L.S. Oliveira, C. Petitjean and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks", In *2016 international joint conference on neural networks (IJCNN),* pp. 2560-2567, IEEE, 2016.

[29] K.H. Yu, C. Zhang, G.J. Berry, R.B. Altman, C. Ré, D.L. Rubin and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features", *Nature communications*, vol.7, p.12474, 2016.

[30] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images", *IEEE transactions on medical imaging*, vol. 35, no.1, pp.119-130, 2015.

[31] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N.N. Shih, J. Tomaszewski, F.A. González and A. Madabhushi, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent", *Scientific reports*, vol. 7, p.46450, 2017.

[32] D. Wang, A. Khosla, R. Gargeya, H. Irshad and A.H. Beck, "Deep learning for identifying metastatic breast cancer", *arXiv preprint arXiv:1606.05718,* 2016.

[33] O. Gevaert, F. Smet, D. Timmerman, Y. Moreau and B. Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks*", Journal of Bioinformatics,* vol.22, pp. e184–e190, 2006.

[34] D. Sun, M. Wang and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp.1–1, 2018.

[35] W. Guo, H. Li, Y. Zhu, L. Lan, S. Yang, K. Drukker, E.A. Morris, E.S. Burnside, G.J. Whitman, M.L. Giger and Y. Ji, "Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data"*, Journal of Medical Imaging,* vol.2, no.4, p.041007, 2015.

[36] F. Hung and H. Chiu, "Cancer subtype prediction from a pathway-level perspective by using a support vector machine based on integrated gene expression and protein network"*, Journal of Computer Methods and Programs in Biomedicine,* vol.141, pp.27–34, 2017.

[37] C. Yu, N. Qin, Z. Pu, C. Song, C. Wang, J. Chen, J. Dai, H. Ma, T. Jiang and Y. Jiang, "Integrating of genomic and transcriptomic profiles for the prognostic assessment of breast cancer"' *Breast cancer research and treatment*, vol. 175, no.3, pp.691-699, 2019.

[38] M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li, Y. Wang and Z. Yang, "Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data", *J Genes,* vol.10, no.3, p.200, 2019.

[39] P. Mobadersany, S. Yousefi, M. Amgad, D. Gutman, J. Barnholtz-Sloan, J. Vega, D. Brat, L. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks", *Journal of Proceedings of the National Academy of Sciences,* vol. 115, no.13, pp. E2970-E2979, 2018.

[40] Z. Wang, M. Li, H. Wang, H. Jiang, Y. Yao, H. Zhang and J. Xin, "Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features", *IEEE Access*, 2019.

[41] "GDC Data portal", 2018. [Online] Available: https://portal.gdc.cancer.gov/. [Accessed: 10 February, 2019].

[42] T. Silva, A. Colaprico, C. Olsen, E. Angelo, G. Bontempi, M. Ceccarelli and H. Noushmehr, "TCGA Workflow Analyze cancer genomics and epigenomics data using Bioconductor packages", 2003.

[43] S. Vijayarani, M.J. Ilamathi and M. Nithya, "Preprocessing techniques for text mining-an overview" *International Journal of Computer Science & Communication Networks*, vol. *5*, no.1, pp.7-16, 2015.

[44] M. Van De Vijver, Y. He, L. Van't Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton and M. Parrish, "A gene-expression signature as a predictor of survival in breast cancer", *Journal of New England Journal of Medicine,* vol. 347, no.25, pp.1999-2009, 2002.

[45] M. Linkert, C. Rueden, C. Allan, J. Burel, W. Moore, A. Patterson, B. Loranger, J. Moore, C. Neves, D. MacDonald and A. Tarkowska, "Metadata matters: access to image data in the real world", *Journal of cell biology,* vol.189, no.5, pp.777-782, 2010.

[46] R. Shaikh, "Feature Selection Techniques in Machine Learning with Python", towardsdatascience.com. [Online] Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Accessed: 23 March, 2019]

[47] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[48] K. Benabdeslem and M. Hindawi, “Efficient semi-supervised feature selection: Constraint, relevance, and redundancy,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1131–1143, 2014.

[49] G. Chandrashekar, F. Sahin, “A survey on feature selection methods”, *Computers and Electrical Engineering,* 2007.

[50] Y. Saeys, I. Inza, P. Larranaga, “A review of Feature Selection techniques in bioinformatics”. Bioinformatics”, *Oxford University press*, 2007.

[51] T. Feng, Fu. Xuezheng, Y. Zhang, A.G. Bourgeois, “A genetic algorithm-based method for feature subset selection”, *Soft Computing*, 2008.

[52] Yu, Kun-Hsing, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." *Nature communications,* vol. 7, p. 12474, 2016.

[53] D. Sun, A. Li, B. Tang, M. Wang, “Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome”, *Computer methods and programs in biomedicine,* vol. 161, pp.45-53, 2018.

[54] Z. Fang, T. Ma, G. Tang, L. Zhu, Q. Yan, T. Wang, J. C. Celedon, W. Chen, and G. C. Tseng, "Bayesian integrative model for multi-omics data with missingness," *Bioinformatics,* vol. 34, no. 22, pp. 3801-3808, 2018.

[55] U.R. Acharya, Y. Hagiwara, V.K. Sudarshan, W.Y. Chan, and K.H. Ng, “Towards precision medicine: from quantitative imaging to radiomics,” *Journal of Zhejiang UniversityScience B*, vol. 19, no. 1, pp. 6-24, 2018.

[56] K. K. Yan, H. Zhao, and H. Pang, “A comparison of graph- and kernel-based –omics data integration algorithms for classifying complex traits,” *BMC Bioinformatics*, vol. 18, no. 1, pp. 1-13, 2017.

[57] P. Lopez-Garcia, A.D. Masegosa, E. Osaba, E. Onieva, and A. Perallos, “Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics,” *Applied Intelligence*, pp.1-16, 2019.

[58] B. Mirza, W. Wang, J. Wang, H. Choi, N.C. Chung, and P. Ping, “Machine Learning and Integrative Analysis of Biomedical Big Data,” *Genes*, vol. 10, no. 2, pp. 1-30, 2019.

[59] “Breast Anatomy”, breast360.org. [Online] Available: https://breast360.org/topics/2017/01/01/breast-anatomy.html/ [Accessed: 23 February, 2019].

[60] “AI vs Machine Learning vs Deep Learning: What's the Difference?”, [Online] Available: https://www.guru99.com/machine-learning-vs-deep-learning.html [Accessed: 23 February, 2019].

List of Publication

[1] Arwinder Dhillon, Dr. Ashima Singh, "Machine Learning in Healthcare Data Analysis: A Survey", *In 3rd International Conference on Data Engineering and Communication Technology ICDECT 2019* [Accepted].

[2] Arwinder Dhillon, Dr. Ashima Singh, "GPELM: An Integrative Analysis for Breast Cancer Survival Prediction", *Computer Methods and Programs in Biomedicine 2019* [Communicated].

thesis

ORIGINALITY REPORT

11%	4%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dongdong Sun, Ao Li, Bo Tang, Minghui Wang. "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome", Computer Methods and Programs in Biomedicine, 2018 Publication	2%
2	www.science.gov Internet Source	<1%
3	ftp.cs.toronto.edu Internet Source	<1%
4	Submitted to National University of Singapore Student Paper	<1%
5	www.ijert.org Internet Source	<1%
6	Submitted to CSU, San Jose State University Student Paper	<1%
7	"Intelligent Computing Theories and Application", Springer Science and Business Media LLC, 2018	<1%