International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 4, July-Aug-2021, ISSN (Online): 2395-566X

Prediction of Breast Cancer using Deep Learning

Asst. Prof. Anandajayam.P, Abhishree.P, Rashmi.C.U, Keerthana.M
Dept. of Computer Science and Engineering,
Manakula Vinayagar Institute of Technology,
Pondicherry, India
anandajayam@gmail.com, pabhishree@gmail.com, rashmicse7@gmail.com, hansukeerthi156@gmail.com

***Abstract*- The recent developments in the information technology have generated a volatile growth of data. In this paper we cover the problem of breast cancer prediction using Machine Learning. We considered two types of datasets, namely, Gene Expression (GE) andDNA Methylation (DM). The main objective of this paper is to predict the breast cancer using Deep Learning. We have developed a deep learning model to detect Breast cancer in CSV files. Benign and Malignant will be the two detections of the model. We use PyCharm as the platformto train and test our datasets.**

***Keywords*- Breastcancer, Geneexpression, DNA methylation, deep learning, CNN, Tensorflow.**

I. INTRODUCTION

Nowadays, the organizations in different sectors are capturing exponentially larger amounts of data than in the past. These data are of different types, spanning over a large family of cases including data from biomedical eld, social networks, sensors, and spatiotemporal stream networks, among others. This huge amount of data requires rethinking and figuring out how to cope in terms of representation, storage, fusion, processing, and visualization.

In the medical field, many data about patients of different diseases are collected every day. Processing these datasets and discovering more valuable knowledge and hidden patterns will improve the medical service and healthcare. Moreover, it will lower the cost of fighting or healing diseases. The fast development of computer science and algorithms has allowed for novel approaches to harness data in order. The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott to discover more insight for competitive advantages, such as classical machine-learning techniques.

Machine learning is considered as one of the fastest growing fields of computer science. Its main concern is enabling computers to learn from input data, usually called training data, and extract knowledge to perform tasks on future data. There are three types of learning: supervised, unsupervised, and reinforcement learning. For each type, several techniques and algorithms exist.

The data samples which are used with machine learning methods are described in terms of features or attributes, which may be of different types and values. The nature of the data decides the type of machine learning techniques to be used in order to obtain valuable information. The analysis of large sets of data is challenging when its aim isto obtain more powerful patterns and information that enable enhanced insight, decision making, and process automation. Unfortunately, the traditional ways of using machine learning algorithms could not cope with the new challenges of big data, especially scalability. Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem.

In today’s society early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumours can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of many researches.

Because of its unique advantages in critical features detection from complex BC datasets, deep learning (DL) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling. This analysis aims to observe whichfeatures are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this we have used deep learning methods to fit a function that can predict the cancer.

II. RELATED WORKS

1. Prediction of Breast Cancer, Comparative Review of Machine:

Breast cancer is type of tumour thatoccurs in the tissues of the breast. It is most common type of cancer found in women around the world and it is among the leading cause of deaths in women. This paper presents the comparative analysis of machine learning, deep learning and data mining techniques being used for the prediction of breast cancer. Many researchers have put their efforts on breast cancer diagnoses and prognoses, every technique has different accuracy rate and it varies for different situations, tools and datasetsbeing used. Our main focus is to comparatively analyse different existing Machine Learning and Data mining techniques in order to findoutthe most appropriate method that will support the large dataset with good accuracy of prediction.

The main purpose of this review is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction and this paper provides the all-necessary information to the beginners who want to analyse the machinelearning algorithms to gain the base of deep learning.

Breast cancer is one of the most lethaland heterogeneous diseases in this present era that causes the death of enormous number of women all over the world. It is the second largest disease that is responsible of women death. There are various machine learning and data mining algorithms that are being used for the prediction of breast cancer.

Finding the most suitable and appropriate algorithm for the prediction of breast cancer is one of the important tasks. Breast cancer is originated through malignant tumours, when the growth ofthe cell got out of control. A lot of fatty and fibrous tissues of the breast start abnormal growth that becomes the cause of breast cancer. The cancer cells spread throughout the tumors that cause different stages of cancer. There are different types of breast cancer which occurs when affected cells and tissues spread throughout the body. Ductal Carcinoma in Situ (DCIS) is type of the breast cancer that occurs when abnormal cells spread outside the breast it is also known as the non-invasive cancer.

The second type is Invasive Ductal Carcinoma (IDC) and it is also known as infiltrative ductalcarcinoma. This type of the cancer occurs whenthe abnormal cells of breast spread over all the breast tissues and IDC cancer is usually found in men. Mixed Tumors Breast Cancer (MTBC) is the third type of breast cancer and it is also known as invasive mammary breast cancer. Abnormal duct cell and lobular cell causes such kind of cancer. Fourth type of cancer is Lobular Breast Cancer (LBC) which occurs inside the lobule. It increases the chances of other invasive cancers. Mucinous Breast Cancer (MBC) is the fifth type that occurs because of invasive ductal cells, it is also known as colloid breast cancer.

It occurs when the abnormal tissues spread around the duct. Inflammatory Breast Cancer (IBC) is last type that causes swelling and reddening of breast. It is a fast growing breast cancer, when the lymph vesselsblock in break cell, this type of cancer starts to appear.

1.1 Techniques Used:

- Nonlinear Algorithms
- Ensemble Algorithm
- Linear Algorithm
- Deep Learning Algorithm
- Stack Sparse Auto Encoder (SSAE),
- Sparse Auto Encoder (SAE) and
- Convolutional Neural Network (CNN) 2.1.2

1.2 Drawbacks:

- The issue of inequality of positive and negative data should be considered by researchers as it can lead to biasness towards positive or negative prediction.
- Another important issue that needed to be solved is imbalanced number of breast cancer images against affected patches for correct diagnosis and prediction of breast cancer.

2. Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features:

A computer-aided diagnosis (CAD) system based on mammograms enables early breast cancer detection, diagnosis, and treatment. However, the accuracy of existing CAD systems remains unsatisfactory. This paper explores a breast CAD method based on feature fusion with Convolutional NeuralNetwork (CNN) deep features.

First, we propose a mass detection method based onCNN deep features and Unsupervised Extreme Learning Machine (US-ELM) clustering. Second, we build a feature set fusing deep features, morphological features, texture features, and density features. Third, an ELM classifier is developed using the fused feature set to classify benign and malignant breast masses. Extensive experiments demonstrate the accuracy and efficiency of our proposed mass detection and breast cancer classificationmethod.

In recent years, deep learning methods,such as the convolutional neural network(CNN), that can extract hierarchical features from image data without the manual selection, which is also called objective features, have been successfully applied witha great improvement on accuracies in many applications, such as image recognition, speech recognition, and natural language processing.

There are some shortcomings in either subjective or objective features. Subjective features ignore the essentialattributes of images, while objective features ignore artificial experience. Therefore, the subjective and objective features are fused so that these features can reflect the essential properties of the image as well as the artificial experience. Meantime, Extreme Learning Machine (ELM) has better classificationeffect on multi-dimensional features than other classifiers including SVM, decision tree, etc., based on our previous research. Thus, we use ELM to classify the extracted breast mass features.

Therefore, in this paper, we proposea novel diagnosis method that merges several deep features.

2.1 Techniques Used:

- Breast image pre-processing
- Mass Detection
- Classify Benign and Malignant massesbased on features fused with CNN feat

2.2 Drawbacks:

- Although the classical diagnosismethod has been commonly used, its accuracy still needs to be improved.
- The quality of the handcrafted feature set directly affects the diagnostic accuracy, and hence an experienced doctor plays a very important role in theprocessof manual feature extraction.

3. Deep Learning Applications in Medical Image Analysis:

The tremendous success of machine learning algorithms at image recognition tasks in recent years intersects with a time of dramatically increased use of electronicmedical records and diagnostic imaging. This review introduces the machine learning algorithms as applied to medical image analysis, focusing on convolutional neural networks, and emphasizing clinical aspects of the field.

The advantage of machine learning inan era of medical big data is that significant hierarchal relationships within the data can be discovered algorithmically without laborious hand-crafting of features. We cover key research areas and applications of medical image classification, localization, detection, segmentation, and registration.

We conclude by discussing research obstacles, emerging trends, and possible future directions. Machine learning algorithms have the potential to be invested deeply in all fields of medicine, from drug discovery to clinical decision making, significantly altering the way medicine is practiced. The success of machine learning algorithms at computer vision tasks in recent years comes at an opportune time whenmedical records are increasingly digitalized. The use of electronic health records (EHR) quadrupled from 11.8% to 39.6% amongst office-based physicians in the US from 2007 to2012.

Medical images are an integral part of a patient's EHR and are currently analysed by human radiologists, who are limited by speed, fatigue, and experience. It takes years and great financial cost to train a qualified radiologist, and some health-care systems outsource radiology reporting to lower cost countries such as India via tele-radiology. A delayed or erroneous diagnosis cause's harm to the patient. Therefore, it is ideal for medical imageanalysisto be carried out by an automated, accurate and efficient machine learning algorithm. Medical image analysis is an active field of research for machine learning, partly because the data is relatively structured and labelled, and it is likely that this will be the area where patients first interact with functioning, practical artificial intelligence systems. This is significant for two reasons. Firstly, in terms of actual patient metrics, medical image analysis is a litmus test as to whether artificial intelligence systems will actually improve patient outcomes and survival. Secondly, it provides a testbed for human AI interaction, of how receptive patients will be towards health altering choices being made, or assisted by anon-human actor.

3.1 Techniques Used:

- History of medical image analysis
- Convolutional neural networks

3.2 Drawback:

- In medical image analysis, the lack of data is two-fold and more acute: there is general lack of publicly available data, and high-qualitylabelled data is even scarcer. Most of the datasets presented in this review involve fewer than 100 patients.
- Yet the situation may not be as dire as it seems, as despite the small training datasets, the papers in this review report relatively satisfactory performance in the various tasks. This is remarkable, as it means that patients can potentially avoid the ionizing radiation from a CT scanner altogether, lowering cost and improving patient safety. NIE also exploited the ability of GANs to generate improved, higher resolution images from native images and reduced the blurriness in the CT images.
- A useful extension of resolution improvement techniques would be applying themto generate MRI images of higher quality.High quality MRI images require high tesla (and correspondingly costlier) MRI scanners. Algorithmically generated high quality MRI images on a lower field strength scanner wouldthus lower healthcare costs.

4. Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks:

Both Wisconsin diagnostic breast cancer (WDBC) database and the Wisconsin breast cancer database (WBCD) are structured datasets described by cytological features. In this paper, we were seeking to identify ways improve the classification performance for each of the datasets based on convolutional neural networks (CNN). However, CNN is designed for unstructured data, especially for image data, which has been proven to be successful in the field of image recognition. A typical CNN may not keep its performance for structured data.

In order to take advantage of CNN to improve the classification performance for structured data, we proposed fully-connected layer first CNN (FCLF-CNN), in which thefully connected layers are embedded before thefirst convolutional layer. We used the fully connected layer as an encoder or an approximator to transfer raw samples into representations with more localities.

In order to get a better performance, we trained four kinds of FCLF-CNNs and made an ensemble FCLF- CNN by integrating them. We then applied it to the WDBC and WBCD datasets and obtained the results by a fivefold cross validation.

The results showed that the FCLF-CNN cans achieve a better classification performance than pure multi-layer perceptrons and pure CNN forboth datasets. The ensemble FCLF-CNN can achieve an accuracy of 99.28%, a sensitivity of 98.65%, and a specificity of 99.57% for WDBC, and an accuracy of 98.71%, a sensitivity of 97.60%, and a specificity of 99.43% for WBCD. The results for both datasets are competitive compared to the resultsof other research.

Convolutional neural networks (CNN) have recently made great success in the field ofimage recognition, objectdetection and image segmentation. In this paper, we were seeking ways to improve the breast cancer classification performance based on CNN. The Wisconsin Diagnostic Breast Cancer (WDBC) database and the Wisconsin Breast Cancer Database (WBCD) are two datasets used for the development of breast cancer automated diagnostic systems. However, they are both structured datasets.

More specifically, each of these samples is described by the existence of cytological features. CNN has been specifically designed for image data. The local connection and multi- layer architecture in CNN can extract multi-level local features in image data, making the CNN outperform other models in the field of image recognition. By breaking down the local structure of the image data, CNN can still work but the performance is inadequate. We illustrate this by empirically exploring the MNIST dataset in Section 2. In order to get better results for both the WBCD and WDBC datasets, we proposed the fully-connected layer first CNN (FCLF CNN), in which the fully connected layers are embedded beforethe first convolutional layer.

We used theselayers as an encoder by setting a softmax loss or an approximator by setting a mean square error (MSE) loss, which can transfer raw samples into representations with more localities. Our experiments demonstrate that the FCLF-CNN can achieve a better classification performance than pure multilayered perceptrons (MLP) and pure CNN.

4.1 Techniques Used

- MLP-CNN
- FCLF-CNN

4.2 Drawbacks

In terms of the complexity of the model, it is true that FCLF-CNN is more complex than the traditional CNN and MLP. But we do not think it is very bad.

- On the one hand, because of the rapid development of hardware, the computational complexity of this level cannot be the main contradiction.
- Second, inthe medical field, in addition to structured data, the proportion of image data is constantlyrising.

III. EXISTING WORK

1. Machine Learning:

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans the ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

1.1 Classification of Machine Learning

Machine learning implementations are classified into three major categories, depending on the nature of the learning “signal” or “response” available to a learning system.

1.2 Supervised Learning

When an algorithm learns from example data and associated target responses that can consist of numeric values or stringlabels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of supervised learning. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules fromthese specific examples.

1.3 Unsupervised Learning

Whereas when an algorithm learns fromplain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlatedvalues. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms. As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the sameclass, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type oflearning.

1.4 Reinforcement Learning

When you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback according to the solution the algorithm proposes comes under the category of Reinforcement learning, which is connected to applications for which the algorithm must make decisions (so the product is prescriptive, not just descriptive, as in unsupervised learning), and the decisions bear consequences. In the human world, it is just like learning by trial and error. Errorshelp you learn because they have a penalty added (cost, loss of time, regret, pain, and so on), teaching algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms. As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the sameclass, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

1.5 Semi-Supervised Learning-

Where an incomplete training signal is given: a training set with some (often many) ofthe target outputs missing. There is a special case of this principle known as Transduction where the entire set of problem instances is known as learning time, except that part of the targets are missing.

1.6 How ML Works-

Gathering past data in any form suitable for processing. The better the quality of data, the more suitable it will be for modelling as givenin fig.1 Data Processing – Sometimes, the data collected is in the raw form and it needs to be preprocessed. Divide the input data into training, cross-validation and test sets. The ratio between the respective sets must be 6:2:2. Building models with suitable algorithms and techniques on thetraining set. Testing our conceptualized modelwith data which was not fed to the model at the time of training and evaluating its performance using metrics such as F1 score, precision and recall.

2. Introduction to Data in Machine Learning

It can be any unprocessed fact, value, text, sound or picture that is not being interpreted and analysed. Data is the most important part ofall Data Analytics, Machine Learning, Artificial Intelligence.

Without data, we can’t train any model and all modern research and automation will go vain. Big Enterprises are spending lots of money just to gather as much certain data as possible.

2.1 Information-

Data that has been interpreted and manipulated and has now some meaningful inference for the users.

2.2 Knowledge

Combination inferred information experiences, learning and insights. Result in awareness or concept building for an individual.

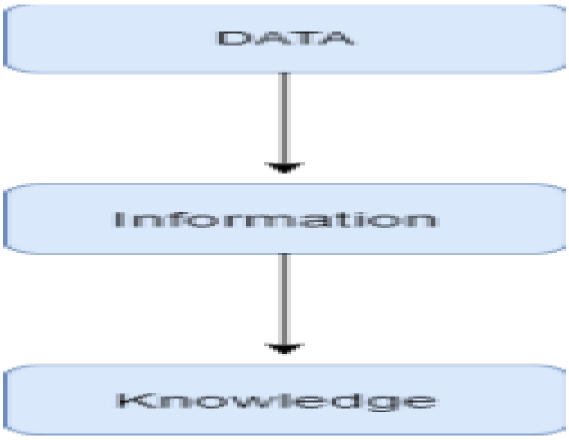

Fig.1. ML working.

3. How Split Data in Machine Learning:

Training Data: The part of data we use to train our model. This is the data which your model actually sees (both input and output) and learn from fig.2

3.1 Validation Data

The part of data which is used to do a frequent evaluation of model, fit on training dataset along with improving involved hyperparameters (initially set parameters before themodel begins learning). This data plays its part when the modelis actuallytraining.

3.2 Testing Data

Once our model is completely trained, testing data provides the unbiased evaluation. When we feed in the inputs of testing data, our model will predict some values (without seeing actual output). After prediction, we evaluate our model by comparing it with actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, setat the time of training.

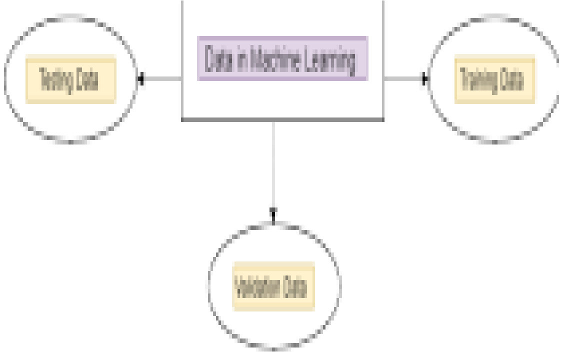

Fig 2. Data splitting.

IV. PROPOSED SYSTEM

Breast Cancer is a second most severecancer among the cancers and unveiled. An automatic disease detection system aids medical staff in disease diagnosis. In this section, the proposed work is explained briefly. The algorithm used and the layers in the model are explained. The goal of this work is to develop a model that is capable of producing the accuracy with respect to the epoch time provided. The epoch time is used to calculate the number of seconds elapsed in the model. The prediction is done from two graph values that is Model Accuracy and Model Loss.

1. Convolutional Neural Network (CNN):
A convolutional Neural Network is a type of Artificial Neural Network that is used in the image recognition and image processing which is designed uniquely to process the pixel data. It contains one or more convolutional layers. It takes ana input image, assign learnable weights and biases to the various objects in the image and be able to differentiate from one from other. The pre-processing for CNN is much lower when compared to other classification algorithms.

The architecture of CNN is analogous to that of the connectivity pattern of the neurons in Human Brain. A CNN is able to successfully capture the spatial and the temporal dependencies in an image through the application of the relevant filters. CNN contains 3 layers. The Convolutional layer, The Pooling layer and The Dense layer. The output layer is also known as fully convoluted layer. The first layer learns basic feature detection filters as edges and corners. The middle layers learn filters that detect the parts of the objects. The last layers have higher representations as they learn to recognise the full objects in various shapes and positions.

1.1 The Convolutional Layer: Convolutional layer is where the filters are applied to the original image or text provided or to other feature maps in a deep CNN. This is where the most of user-specified parameters are in the network.
The actions of the Convolutional layer are

- Apply filters to extract features
- The filters are composed of small kernels are learned
- One bias per filter
- Apply activation function on every valueof the feature map
- The parameters of Convolutional layer are the number and size of kernels
- Activation function
- Padding
- Regularization type and value

The input and output format of Convolutionallayer are

- INPUT: 3D cube is the previous set of featuremaps
- OUTPUT: 3D cube and one 2D map per filter.

1.2 The Pooling Layer: The Pooling layers are similar to the convolutional layers, they perform a specific function such as max pooling, that takes the maximum value in a certain filter region, or average pooling that takes the average value in the filter region. These are the typically used reduce dimensionality of the network.
The actions taken by the Pooling layer are

- To reduce dimensionality
- Extract the maximum of average of a region
- The sliding window approach

The parameters of the pooling layer are

- Size of the window
- The input and output of the pooling layer are
- INPUT: 3D cube and the previous set of featuremaps
- OUTPUT: 3D cube, one 2D map per filter andthe reduced spatial dimensions

1.3 The Fully Connected Layer: The fully connected layers are placed before the classification output of a CNN and are used to flatten the results before classification. It is similar to the output layer of an MLP.

The action taken by the fully connected layer is

- Aggregate information from the final feature maps
- Generates the final classification

The parameters of the fully connected layer are

- The number of nodes
- Activation function

The input and output of the fully connected layer are

- INPUT: Flattened 3D cube and the previous setof feature maps
- OUTPUT: 3D cube and one 2D map per filter

The role of CNN in this model is that it divides the given csv input into layers. It layers the data into a three dimensional data.

2. Model Evaluation:
The result of the model is attained using model accuracy and model loss of the developed model.
2.1 Model Accuracy: The model accuracy graph is obtainedfrom the epoch time and the accuracy of thetraining and the testing datasets. The training of the model can be done until the validation accuracy is greater than the training accuracy. This proves that the model has not still attained the overfitting. The result of the model is described in figure.3

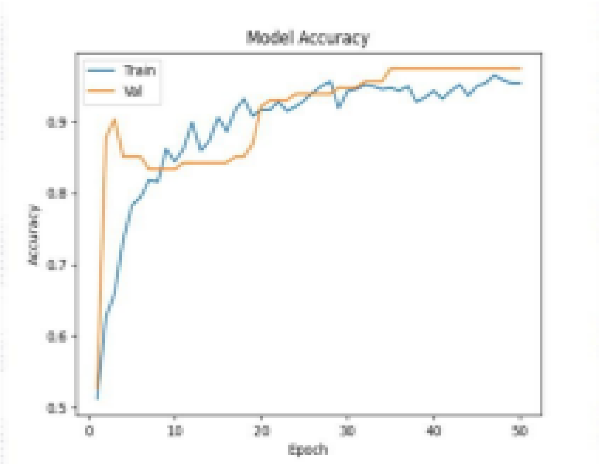

Fig 3. Model Accuracy.

2.2 Model Loss:

The model loss graph is obtained from the epoch time, the validation loss and training loss. The training of the model can be done until the validation loss goes greater than the training loss. The result of the model loss is described in the figure.4

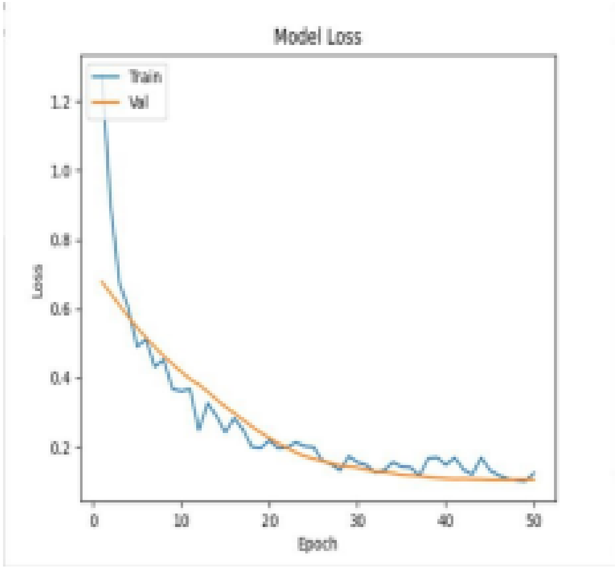

Fig 4. Model Loss.

V. CONCLUSION

Various deep-learning techniques can be used for the prediction of breast cancer. But, thereal challenge is to come about with a efficient medical data classifiers which is accurate and computationally. In this study, we have focused on analysing the dataset of breast cancer patients and predicting their Gene type. We used Pycharm as the platform to find accuracy and have employed CNN algorithm for getting acurate result of the prediction. The must note factor in our project is that we have tried to get better accuracy rate by using clean dataset and using deep learning algorithm which as not been done yet.

The main aim of our project is to make prediction easy and accurate in less time. The predication rate of our CNN is 97% for which model accuracy and model loss is given. Futhermore, this project can be made even better by increasing the datasets.

REFERENCE

[1] K. P. Murphy, Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2012.

[2] M. Guller, Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis. Berkeley, CA, USA: Apress, 2015.

[3] International Agency for Research on Cancer (Iarc) and World Health Organization (Who). Globocan 2018: Age Standardized (World) Incidence and Mortality Rates, Breast. Accessed: Sep. 1, 2018. [Online].

[4] (2016). DNA Deoxyribonucleic Acid. [Online]. Available:http HYPERLINK "http://www.myvmc.com/anatomy/d%2 0"://www.myvmc.com/anatomy/dna- deoxyribonucleic-acid/

[5] Y. Lu and J. Han, ``Cancer classification using gene expression data," Inf.Syst, vol. 28, no. 4, pp. 243268, 2003.

[6] M. M. Babu, ``Introduction to microarray data analysis," Comput.Genomics, Theory Appl., vol. 17, no. 6, p. 249, 2004. (2018).

[7] Spark Programming Guide Spark2.0.1 Documentation. Accessed: Oct. 15, 2018. [Online]. Available: https://spark.apache.org/docs/2.0.1/programmming-guide.html

[8] (2018). Weka 3 Data Mining with Open Source Machine Learning Software in Java. [Online] Available:https://www.cs. waikato.ac.nz/ml/weka

[9] A. Kowalczyk, ``Support vector machines succinctly," Syncfusion, Inc., Morrisville, NC, USA, 2017.

[10] J. R. Quinlan, ``Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp.106, 1986.

[11] A.Lia and M.Wiener, ``Classification and regression by randomforest,"R News, vol. 2, no. 3, pp. 180222, 2002.

[12] V. Chaurasia and S. Pal, ``A novel approach for breast cancer detection using data mining techniques," Int. J. Innov. Res. Comput. Commun Eng., vol. 329, no. 1, pp. 23209801, 2014.

[13] B. Zheng, S. W. Yoon, and S. S. Lam,``Breast cancer diagnosis based on feature extraction using a hybrid of K- means and support vector machine algorithms," Expert Syst. Appl., vol. 41, pp. 14761482, Mar. 2014.

[14] C. Park, J. Ahn, H. Kim, and S. Park``Integrative genenetwork construction to analyze cancer recurrence using semi- supervised learning," PLoSONE, vol. 9, no. 1, 2014, Art. no. e86309.

[15] R. B. Ray, M. Kumar, and S. K. Rath,``Fast in-memory cluster computing of sizeable microarray using spark," in Proc. Int. Conf. Recent Trends Inf.Technol. (ICRTIT), Chennai, India, 2016, pp. 16. (2018).

[16] National Center for Biotechnology Information. [Online]. Available:https://www.ncbi.nlm.nih.gov.

[17] Y. V. Lokeswari and S.G. Jacob, ``Prediction of child tumors from microarraygene expression data through parallel geneselection and classification on spark," in Computational Intelligence in Data Mining.Singapore: Springer, 2017, pp. 651661.