JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AUTOMATED BREAST MASS CLASSIFICATION SYSTEM USING DEEP LEARNING AND ENSEMBLE LEARNING IN DIGITAL MAMMOGRAM

KarthikaK[1],SowmiyaP[1],TamilselvanSP[1], SudhaG[2]

[1] UG Student, Department of Medical Electronics, Muthayammal Engineering College, Namakkal,Tamilnadu.

[2] Professor, Department of BioMedical Engineering, Muthayammal Engineering College, Namakkal,Tamilnadu.

KEYWORDS

Deep Learning
CNN
Machine Learning
Mammogram
Breast Cancer

ABSTRACT

Breast cancer is one of the most highly invasive malignancies tumors that occurs in women and rarely in men . It is examined the worst cancer after lung cancer due to the higher death rate in women. Every year number of death is increasing extremely because of breast cancer. It is the most often type of all cancers and the major cause of death in women worldwide. Any development for prediction and diagnosis of cancer disease is most important for a healthy life. Accordingly, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Breast mass classification systems are implemented using deep learning technologies such as a Convolutional Neural Network (CNN). CNN based systems have attained higher performance than the machine learning-based systems in the classification task of mammography images. The objective of this project is to collect and train the dataset, to predict the survival of patients with breast cancer. Based on the results, we can compare the performance of both ML and DL process (we implemented). To get the best model which is suitable for breast cancer detection.

1. INTRODUCTION

Breast cancer is a classification of cancer that occurs in the breast tissue. It's important to understand that most breast lumps are benign and malignant. By using ML &DL techniques, find out the stage whether it is benign or malignant. The systematic study of statistical tools and algorithms that computer systems use to perform a specific work without using explicit instructions, depending on patterns is Known as Machine Learning. It is seen as a subdivision of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, in order to make predictions or decisions without being distinct programmed to perform the work. Deep learning depends on the set of machine learning techniques that model high- level abstractions in data with multiple non-linear processes. Deep learning method operates on the artificial neural network framework. Such ANNs are constantly using neural networks, and the performance of training can be enhanced by having increased the quantity of data. Efficiency relies on larger data quantities. The training method is called deep so because amount of neural network rates rises. The functioning of the deep learning experience depends solely on two stages, the training phase and the testing phase.

2. LITERATURE SURVEY

[1]In this paper, the various technologies of data mining (DM) models for forecast of heart disease are discussed. Data mining plays an important role in building an intelligent model for medical systems to detect heart disease (HD) using data sets of the patients, which involves risk factor associated with heart disease. Medical practitioners can help the patients by predicting the heart disease before occurring. The large data available from medical diagnosis is analyzed by using data mining tools and useful information known as knowledge is extracted.

[2]A novel ensemble learning approach "BBS method" which stands for Bagging, Boosting and Stacking with appropriate base classifiers for the classification of the five UCI datasets taken from the field of Bioinformatics. Experiments are conducted gives better accuracy with lower root mean square error rate using the technique of learning method is more suitable in handling the classification problem in the bioinformatics domain. Such approaches can be efficiently used in related real-life scenarios of classification domain.

[3].Enormous data growth in multiple domains has posed a great challenge for data processing and analysis techniques. In particular, the traditional record maintenance strategy has been replaced in the healthcare system. It is vital to develop a model that is able to handle the huge amount of e healthcare data efficiently. In this paper, the challenging tasks of selecting critical features from the enormous set of available features and diagnosing heart disease are carried out. Feature selection is one of the most widely used pre-processing steps in classification problems.

[4]Crime is a foremost problem where the top priority has been concerned by individual, the community and government. It investigates a number of data mining algorithms and ensemble learning which are applied on crime data mining Crime forecasting is a way of trying to mining out and decreasing the upcoming crimes by forecasting the future crime that will occur.

[5].In this work, Chronic kidney disease is a fatal illness of kidney which can be prevented with early correct predictions and proper precautions. Data mining of the information collected from previously diagnosed patients opened up a new phase of medical advancement. However, specific techniques must be executed to accomplish better consequence. In this manuscript the capability of the classification of Support Vector Machine, Decision tree, Naïve Bayes and K-Nearest Neighbor algorithm, in analyzing the Chronic Kidney Disease dataset collected from UCI repository, was investigated to predict the presence of kidney disease. Dataset has been analyzed in terms of accuracy, Root Mean Squared Error, Mean Absolute Error and Receiver Operating Characteristic curve. In the present study, Decision tree shows promising results.

3. Related Work

In existing work, they used the Convolutional Neural Networks to detect & classify the breast cancer. But they selected the priority & non priority features based on heat map correlation using Seaborn library. Correlation feature values may confuse the users & researchers, because all values are in decimal with negative values also that time, there is possibly to missed some priority features from dataset. They use the data split with 75% -25% for training & testing.

3.1 CNN Architecture

Convolutional neural networks (CNN) are a class of artificial neural networks and one of the

most common deep learning architectures for image recognition tasks. The model takes an input image, performs a series of convolutional and pooling layers, followed by fully connected layers to execute the output.

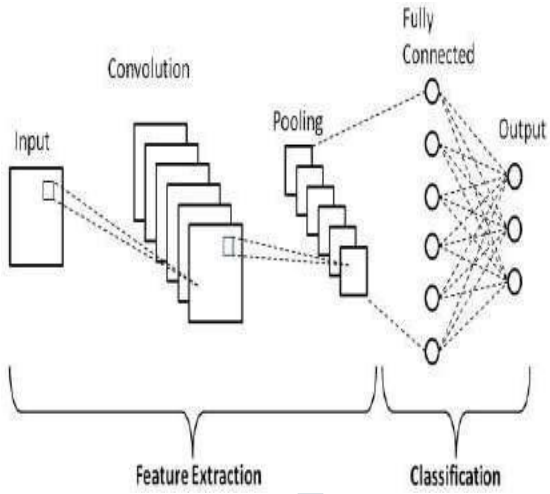


Fig.3.1 CNN Architecture

3.1.1 Convolutional layers

The term convolution means the mathematical combination of two functions to form a third function. When that happens, two sets of information are merged. In the context of CNNs, a convolutional layer is applied to the input data to then produce a feature map. The training dataset is used to train the model with. in case of neural network, the model it has weight and biases. The validation dataset is what the model uses for evaluation after every set of predictions. The test dataset is used to evaluate the model after it has been completely trained.

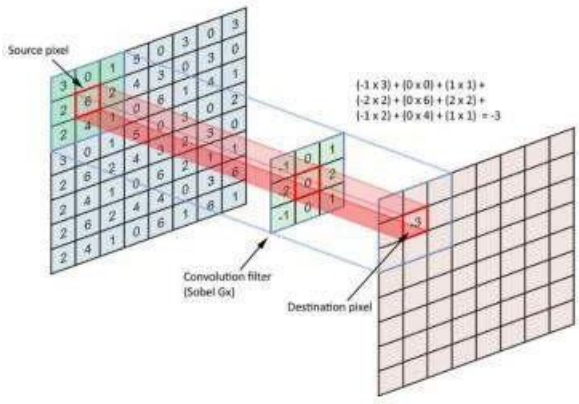


Fig.3.1.1.Convolutional Layers

There are two other important concepts in convolutional layers: strides and padding. Strides are the number of pixel a kernel or a filter slides over the input matrix. Padding is what is used when the filter does not fit the input matrix. There are two types of padding: valid padding, when the border pixels of the input matrix are discarded; and zero or same padding, when zeros are added to the borders so that the filter fits the input matrix.

3.1.2 Pooling layers

Pooling layers are responsible for reducing the dimensionality of feature maps, specifically the height and width, preserving the depth. Doing so is beneficial because it decreases the required computational power to process the data, while extracting the dominant features in feature maps. There are two types of pooling layers: max pooling and average pooling.

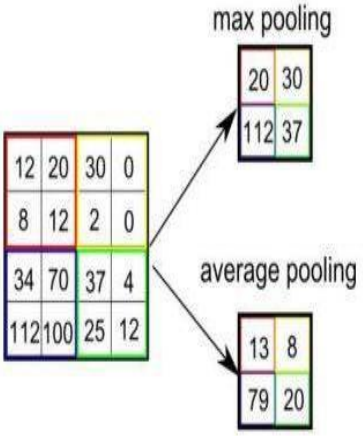


Fig.3.1.2. Pooling Layers

Max pooling outputs the maximum value of the elements in the portion of the image covered by the filter, while average pooling returns the average value. Max pooling is better at extracting effective features and therefore considered more perfomant.

3.1.3 Fully connected layers

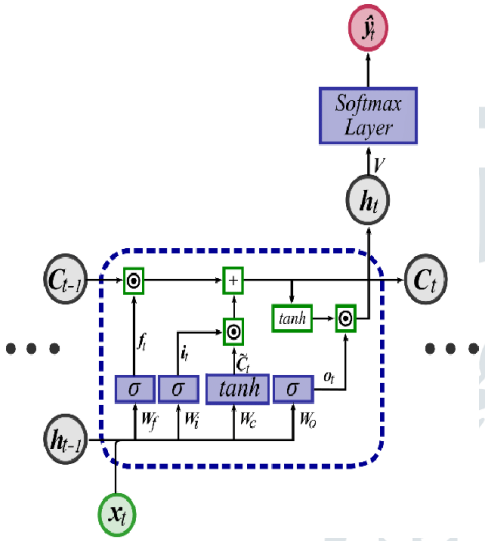
Fully connected layers are where classification actually happens. The input matrix is flattened into a column vector and is fed into a set of fully connected layers which are the same as the fully connected ANN architecture . Each fully connected layer (called Dense layer) is passed through an activation function , but the output Dense layer is passed through Softmax. In the Softmax multiclass classification, the loss function used is Cross Entropy.

The output of the Softmax function is an N-dimensional vector, where N is the number of classes the CNN has to choose from. Every number in this N dimensional vector represents the probability that the image belongs to each certain class. For example, if the output vector is [0 .1 .1 .75 0 0 0 0 0

.05], then there is a 10% probability that this image belongs to the class 2, 10% probability that it belongs to the class 3, 75% probability that this image belongs to the class 4, and 5% probability that it belongs to the class.

3.2 Artificial Neural Networks

An artificial neural network (ANN) is a set of layers of neurons (in this context they are called units or nodes). In the case of a fully connected ANN, each unit in a layer is connected to each unit in the next layer .



$$\tanh(x) = 2\,sigmoid(2x) - 1$$

Fig.3.2 Artificial Neural Network

There is an input layer, where the network takes all the information needed, in this case the images to categorize. Between the input layer and the output layer are hidden layers. Each hidden layer is used to detect a different set of features in an image, from less to more detailed. For example, the first hidden layer detects edges and lines, the second layer detects shapes, the third layer detects certain image elements, for example a face or a wheel.

The output layer is where the network makes predictions. The predicted image categories are compared to the labels provided by humans. If they are incorrect, the network uses a technique called back transmission to correct its learning, so it can make guess more correctly in the next iteration.

3.3 RNN-LSTM MODEL

A recurrent neural network is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to reveal temporal dynamic behavior. Derived from feed forward neural networks, RNNs can use their internal state to process variable length sequences of inputs. This makes them applicable to tasks such as connected handwriting recognition or speech recognition.

$$f(x) = tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

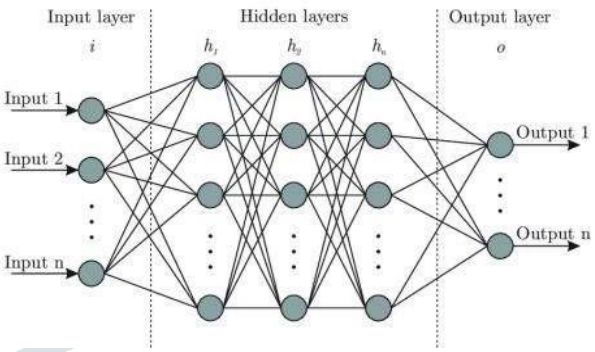


Fig.3.3 RNN-LSTM Model

LSTM can learn tasks that require memories of events that happened thousands or even millions of discrete time steps earlier. Problem- specific LSTM-like topologies can be evolved.

Many applications use stacks of LSTM RNNs and train them by Connectionist Temporal Classification (CTC) to find an RNN weight matrix that maximizes the probability of the label sequences in a training set, given the corresponding input sequences. CTC achieves both

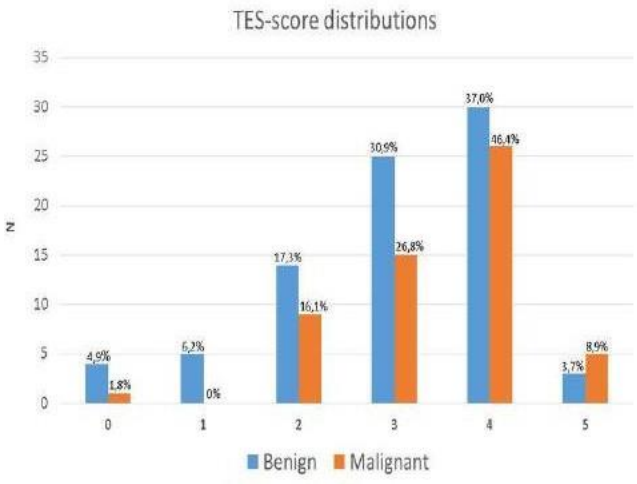
alignment and recognition. LSTM can learn to recognize context-sensitive languages unlike previous models based on hidden Markov models (HMM) and similar concepts.

3.4 MODEL TRAINING AND RESULTS

Hardware and software used The thesis uses free TPU/GPUs from Google Colab (Colaboratory). The deep learning framework used is Tensor Flow with Keras API.

```
X_train, X_test, y_train, y_test = train_te st_split(Xa_1, Target_Class, random_sta te=0 , test_size=0.30,stratify=Target_Cla ss)

display("for training")

print(X_train.shape,

y_train.shape) display("for

testing") print(X_test.shape,

y_test.shape)
```

3.5 Target Class



3.6 Training, Valid and test

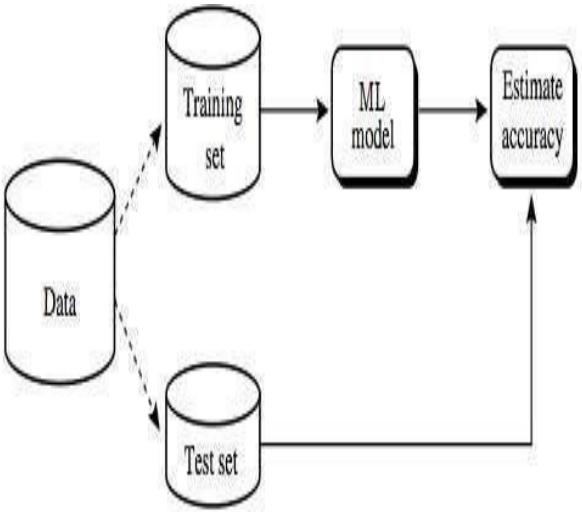


Fig. 3.6 Training and Testing

The train dataset is used to train the model with. In the case of neural networks, the model learns its weights and biases. The validation dataset is what the model uses for evaluation after every set of predictions. The test dataset is used to evaluate the model after it has been completely trained.

After that, collection of ML performance when compared with 3 different custom CNN models. To identify the best classifier to detect & classify the phishing website. The complete implementation can be done through Google Colab (Python-Jupyter Notebook).

4. Experiment methods and results

Initially we have 33 features are taken up for the machine learning classification. Before initiate the machine learning process, we are going to use visualize the data using plotly & Seaborn library. After that we used the classifiers like Support vector, K Nearest neighbour, decision tree classifier, Random forest. Before passing the dataset to the classifier, entire data can be split into 2 parts as training and testing.80% allocated for training & 20% allocated for testing. After the data split, training data is applied to the machine learning classifiers. Based on the training, a test data is taken up for validation or prediction i.e. in order to find the performance of the classifiers.

There are several kinds of machine learning classification methods. The main two of them are supervised and unsupervised instruction. Other than those mentioned above, are reinforcement learning, recommended systems, etc. The concepts of issues that can be discussed by machine learning methods are regression, classification and clustering. The supervised machine learning job is undergoing two major phases, i.e. the training and the testing period. During the training period, the graph is built and evaluated in the same way during testing stage. The possibility of using machine learning for decision-making in a variety of fields has a critical part to play in people's advancement. There are several various machine learning techniques, and the 3 main types under which they can be categorized are supervised, unsupervised, and reinforced learning.

4.1 Training Models

Generally, machine learning models require a lot of data in order for them to perform well. Usually, when training a machine learning model, one needs to collect a large, representative sample of data from a training set. Data from the training set can be as varied as a corpus of text, a collection of images, and data collected from individual users of a service. Over fitting is something to watch out for when training a machine learning model.

4.2 Support vector machines

Support vector machines also known as support vector networks, are a set of related supervised learning method used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one form or the other. An SVM training algorithm is a non-probabilistic, binary, linear classifier, although methods such as Platt scaling exist to use SVM in numerical classification setting.

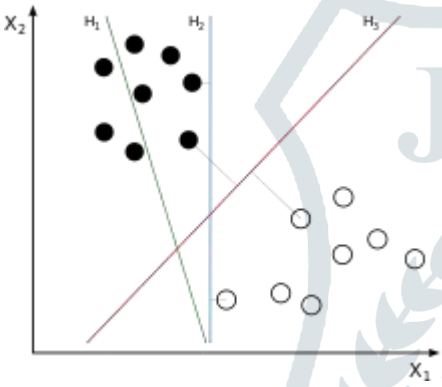


Fig.4.2 Support Vector Machine

4.3 Decision trees

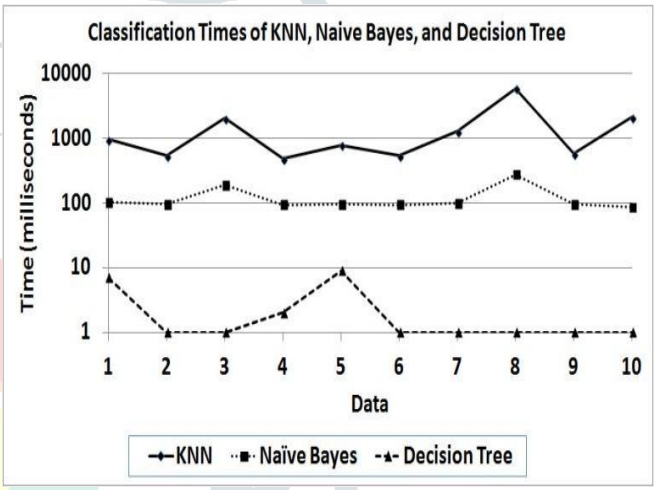
In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data, but the resulting classification tree can be an input for decision making.

4.4 k-Nearest-Neighbor

k-Nearest-Neighbor classification is a machine learning algorithm that localizes a group of k objects in a training case that has the closest proximity to the test object, and then assigns a label derived from the prevalence of a class in the closest proximity. Three important elements are needed for this algorithm, a group of labeled objects; a proximity metric; and the number k of nearest neighbors.

4.5 Naive Bayes

A Naive Bayes classifier may be a probabilistic machine learning model that is used for the classification tasks. The crux of the classifier is predicated on the Bayes theorem. Using Bayes theorem, we will find the probability of an event, as long as B has appeared. Here, B is that the evidence and A is that the hypothesis. the idea made here is that the features are independent, that is the presence of one particular feature that does not affect the opposite. Hence it's called naive**.**



4.6 Regression analysis

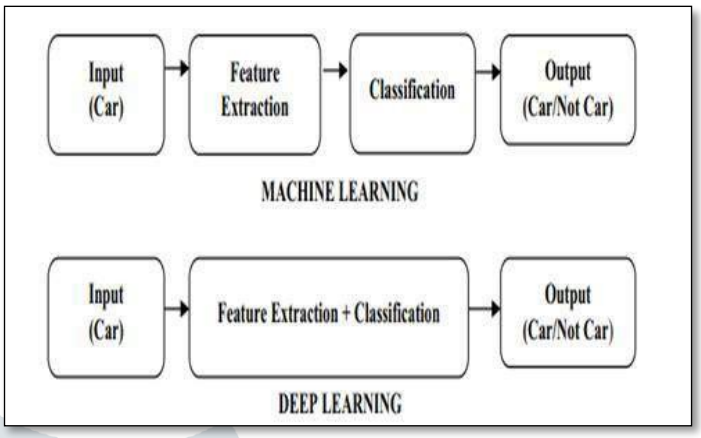
Regression analysis encompasses a large variety of statistical methods to evaluate the relationship between input variables and their associated features. Its most common form is linear regression, where a single line is drawn to best fit the given data according to a mathematical criterion such as ordinary least squares

4.7 Random forest

Random forests also known as random decision forests generate a large number of trees that achieve their output through ensemble learning methods for classification, regression. Bagging and feature randomness are the features it uses to establish those trees. The random forest has an advantage over the decision tree which, is that it does not over fit the data.

4.9 Differences between ML &DL

After applying the Random forest feature selection, less priority features are removed. Only 15 features (which are highly important). Then priority features are again applied to the machine learning classifiers to get the performance.



.

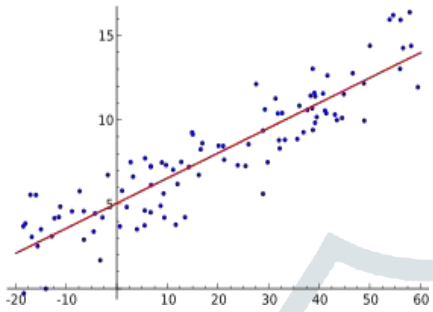


Fig.4.7. Random Forest

4.8 Classifier Results

	model	feature_count	acc	prc	rec	f1
0	rfc_model_1	30	0.935673	0.935673	0.935673	0.935673
1	rfc_model_2	30	0.947368	0.947368	0.947368	0.947368
2	dtc_model_1	30	0.906433	0.906433	0.906433	0.906433
3	nbc_model_1	30	0.912281	0.912281	0.912281	0.912281
4	svc_model_1	30	0.923977	0.923977	0.923977	0.923977
5	knc_model_1	30	0.935673	0.935673	0.935673	0.935673
6	rfc_model_1a	15	0.959064	0.959064	0.959064	0.959064
7	rfc_model_2a	15	0.964912	0.964912	0.964912	0.964912
8	dtc_model_1a	15	0.953216	0.953216	0.953216	0.953216
9	nbc_model_1a	15	0.947368	0.947368	0.947368	0.947368
10	svc_model_1a	15	0.923977	0.923977	0.923977	0.923977

5. Software Requirements

5.1 PYTHON AND ITS FEATURES

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems.

Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is considered to be highly readable.

Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently whereas other languages use punctuation, and with as fewer syntactical constructions than other languages.

•It supports functional and structured programming methods as well as OOP.

•It can be used as a scripting language or can be compiled to byte-code for building large applications.

•It provides very high-level dynamic data types and supports dynamic type checking.

•It supports automatic garbage collection.

5.2 Libraries & Packages in Python

5.2.1 Keras:

Keras is an open-source library that provides a Python interface for artificial neural networks. Keras acts as an combine for the TensorFlow library. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

5.2.2 TensorFlow

TensorFlow was developed by the Google Brain team for internal Google use. TensorFlow is a free and open-source software library for machine learning. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. Tensorflow is a symbolic math library based on dataflow and variation programming. It is used for both research and production at Google.

5.2.3 Plotly

Plotly is a technical computing company headquartered in Montreal, Quebec, that develops online data analytics and visualization tools

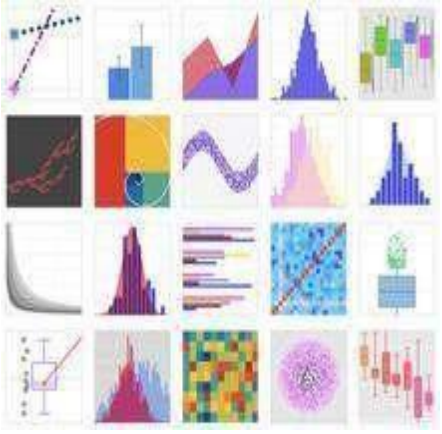


Fig.5.2.3 Plotly

Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

5.2.4 Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI tool kits like Tkinter, wxPython, Qt.

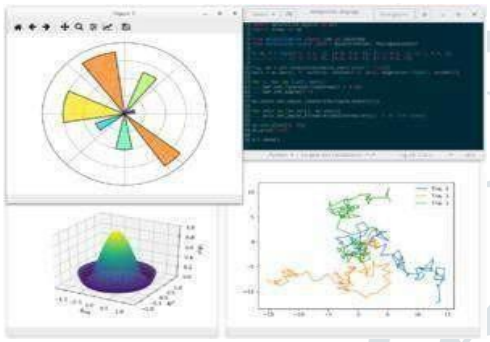


Fig.5.2.4 Matplotlib

5.2.5 Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high- level interface for drawing attractive and informative statistical graphics

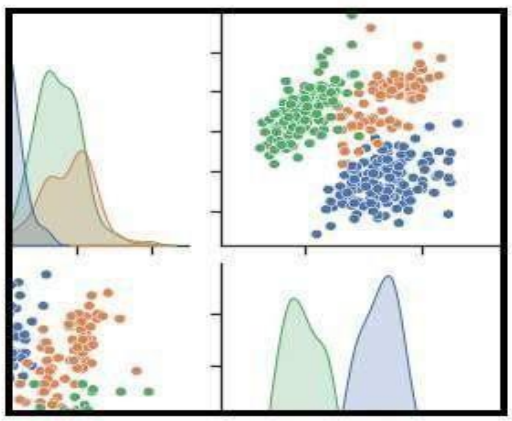


Fig.5.2.5 Seaborn

5.2.6 Pandas

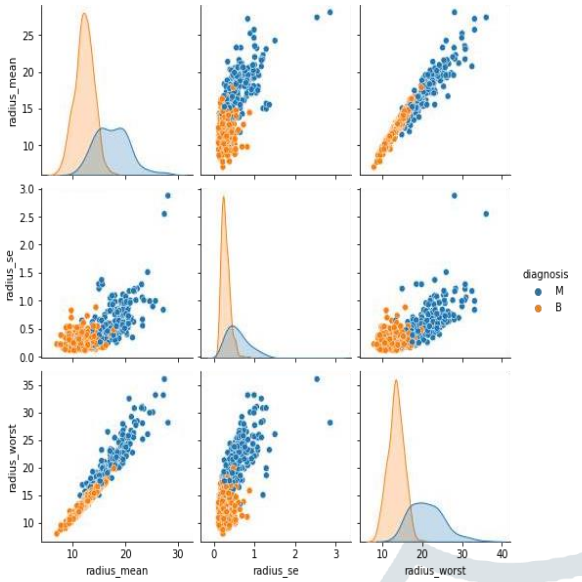
Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. Pandas is a Python data analysis library and is used essentially for data manipulation and analysis. It comes into play before the dataset is prepared for training.

Pandas is one of the tools in Machine Learning which is used for data cleaning and analysis. It has features which are used for analysing, cleaning, transforming and visualizing from data. Pandas is mainly used for data analysis. Pandas allows various data manipulation operations such as merging , reshaping , selecting , as well as data cleaning, and data wrangling features.

5.2.7 Numpy

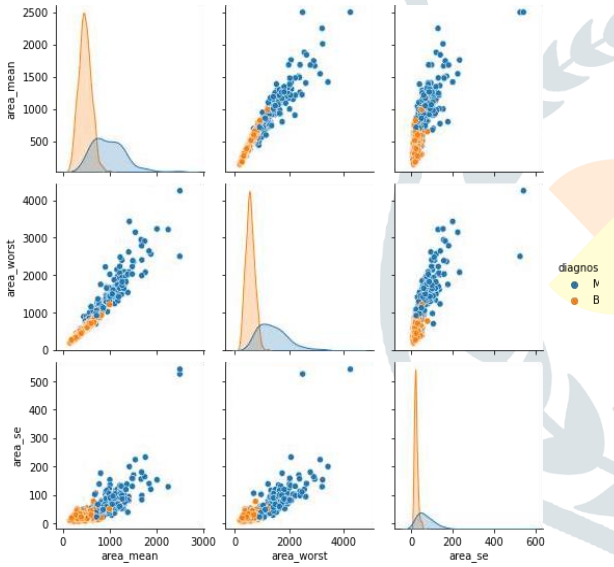
It is a library for the Python programming language, adding assist for large, multi- dimensional arrays and matrices, along with a large collection of high-level mathematical functions to do on these arrays. NumPy by incorporating features of the competing Num array into Numeric, with considerable modifications. NumPy is open-source software and has many developers

6. Data Visualization



sns.pairplot(y1[['area_mean','area_worst' ,'area_se','diagnosis']],hue='diagnosis')

7.



Conclusion and Results

Breast cancer if identified at an early stage will help save lives of thousands of women or even men. These projects help those who are all affected by breast cancer and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for our project proposed by us. In our work, by the help of RFE feature selection Process to get the important features, applied to the classifier(Both ML &DL). Based on our set of machine learning and custom CNN models implementation.

Using ML &DL algorithms we will be able to classify and predict the cancer into being or malignant and find out which is best suitable for prediction. These algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes. We used the programming language of python for model accuracy. In python, lot of libraries and packages are there. We have implemented in seaborn library, it is highly effective for data visualization . Even though machine learning models getting %accuracy, But our customized CNN model-1 with SGD optimizer getting % accuracy, slight better accuracy when compared with dtc_model. Therefore, ML/DL model is best for our breast mass classification approach.

8. Future Work

Right now, we have less than 1000 records taken up for this prediction. In future, we can gather a huge volume of patients data, it will be processed & applied towards our customized Deep CNN model (more features added).

9. References

[1] Lekha, S., & Suchetha, M. (2017). A novel 1-D convolution neural network with SVM architecture for real- time detection applications. IEEE Sensors Journal, 18(2), 724-731.

[2] Jayasree, T., Bobby, M., & Muttan, S. (2015). Sensor data classification for renal dysfunction patients using support vector machine. Journal of Medical and Biological Engineering, 35(6), 759- 764.

[3] I.H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[4] F.Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn.Res.,vol. 12, no. Oct, pp. 2825– 2830, 2011.

[5] Benjamin EJ, Virani SS, Callaway CW, et al. Heart disease and stroke statistics— 2018 update: a report from the American Heart Association. Circulation. 2018 Mar 20;137(12):e67–492.

[6] Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technol Health Care. 2016 Jan27;24(1):31–42.

[7] Witten IH, Frank E, Hall MA. The WEKA workbench. Online appendix for "Data mining: practical machine learning tools and techniques." 4th ed. Morgan Kaufmann;2016.

[8] Anand RS, Stey P, Predicting mortality in diabetic ICU patients using machine learning and severity indices. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci.2018;2017:310–9.

[9] A. Rampun, B. W. Scotney, P. J. Morrow, and H. Wang, ''Breast mass classification in mammograms using ensemble convolutional neural networks,'' in Proc. IEEE 20th Int. Conf. E-Health Netw., Appl. Services (Healthcom), Sep. 2018, pp.1–6.

[10] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, ''A survey of sparse representation: Algorithms and applications,'' IEEE Access, vol. 3, pp. 490–530,2015.

[11] "A Comparative Study of Ensemble Learning Methods for Classification in Bioinformatics " by Aayushi Verma, Shikha Mehta in 2017 7th International Conference on 2017 Jan 12 (pp. 155-158).IEEE.

[12] "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease "by T. Vivekananda, N. Ch Sriman Narayana Iyengar in 0010- 4825/2017 ElsevierLtd**.**

[13] "Survey Paper on Crime Prediction using Ensemble Approach" by Ayisheshim Almaw, Kalyani Kadam in International Journal of Pure and Applied Mathematics2018.

[14] Nusrat Tazin, Shahed Anzarus Sabab , Muhammed Tawfiq Chowdhury in 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)

[15] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, et al. Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nat Med* (2019) 25(8):1301–9. doi: 10.1038/s41591-019-0508-1

Welcome to Python.org [Internet]. Python.org. [cited 2018 Aug 5]. Available from: https://www.python.org/