International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Predictive Analysis of Breast Cancer by Visualizing through Topological Data Analysis

Swapnil Roy[1], Shivam Kalhans[2], Merin Meleet[3], Dr. Rajashekara Murthy S[4]

[12]*Student, Dept. of Information Science and Engineering, R.V. College of Engineering, Karnataka, India*
[3]*Assistant Professor, Dept. of Information Science and Engineering, R.V. College of Engineering, Karnataka, India*
[4]*Associate Professor, Dept. of Information Science and Engineering, R.V. College of Engineering, Karnataka, India*

Abstract - *Women in India are faced with major fatality which includes respiratory or orthopaedic problems but some face major illness in the form of breast cancer which varies for every women, almost fifty percent of the middle age women are diagnosed with this deadly disease. Breast Cancer are actually detected by looking for lumps like tumour present in women's breasts. The cells found in these lumps may be uneven or unstructured leading to malignant tumours which can cancerous and have to be treated immediately. But there might be lumps with cells of even size and having no structural difference amongst each other leading to benign tumours which are totally non-cancerous. This paper focuses on building a machine learning algorithm to predicting the two different tumours whether benign or malignant and visualize the features of both the tumours through Topological Data Analysis. The most important features used for separating the two classifications of tumours will then be visualized through bar graph plotting.*

Key Words: **Breast Cancer, KMapper, t-SNE, GaussianNB, RandomBoosting Classifier, Logistic Regression, Topological Data Analysis(TDA)**

1. INTRODUCTION

Breast cancer is very frequent form of cancer in females. The cancer affects almost all the vital organs of the body if found chronic and eventually leads to death in some cases. The cases for breast cancer during the period 2010-14 showed huge variations with changing locations. The results generally showed 50% and even above it. There are no preventive measures taken now to prevent the spread of cancer. But researches prove that the early detection and diagnosis of tumours present can cure the disease with correct treatment and improve the chances of survivability.

The symptoms of the disease are not clear in the early stages which might be the sole reason for the treatment getting delayed and increasing number of casualties. Therefore, the medical practitioners have advised all women above the age of forty to consult a doctor and should get a mammogram almost every year to see the symptoms if there are any. A mammogram actually is an X-ray scanning for the tumours present in a breast. However sometimes the symptoms for a benign tumour and malignant tumour might actually look the same. This causes panic, if the tumours are not clearly differentiated. The benign tumours are usually the non-cancerous growth of lumps in the body which actually might just be a cyst with its cells all of the same and even texture. However, a malignant tumour is a cancerous outgrowth which might be deadly in certain conditions with a fast reproductive nature of the cells. The cells might be uneven and might have completely different properties when compared to the neighbouring cells. The presence of these cancerous tumours are prevalent in urban areas and have been increasing globally over the years. An early detection of the tumours might help the treatment and diagnosis getting started at a very premature stage and help in curing the disease through regular chemotherapy sessions.

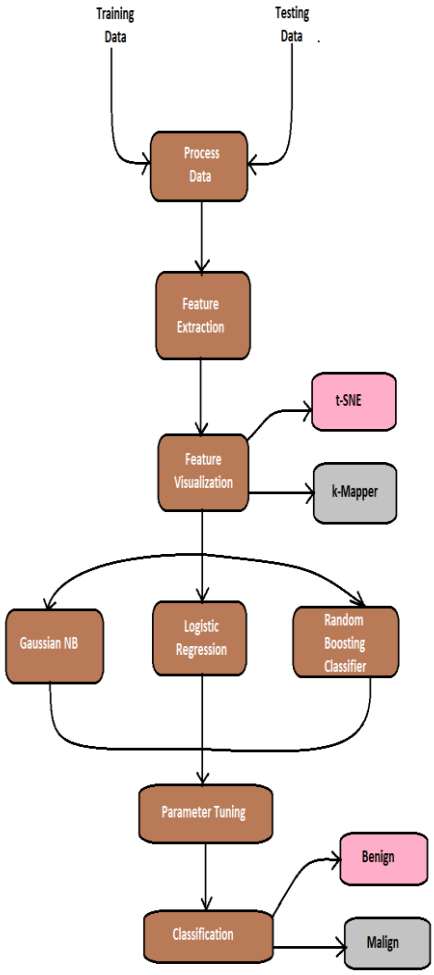

Figure-1: Proposed design describing the flow of the model

2. RELATED WORK

The principles of [1] are actually very different from the recent works on how to combat the deadly cancer, it takes a completely different route altogether. The research of prognosis is focused on three measures which includes: 1) predicting all the risk assessments that might lead to the disease 2) predicting what are the chances of tumor to grow again and 3) chances of the patient surviving from the disease. However, going by the current scenario of latest technology [3] finds a deep learning method by applying methods of CNN for segregating X-rays which overwhelmingly performed much better than any previous models. The results were presented on a digitized film where only one model achieves an image AUC score of 0.88 whereas on improving the model to four averaging model gave the AUC score of 0.91. [4] gave as an insight on Artificial Intelligence techniques that is capable of surpassing human experts in breast cancer prediction. [2] presented a rather or basic technique that compares algorithms like Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes. The results obtained were quite interesting and have quite a good metric for further treatments and detection. [5] and [6] incorporated the technique for a hybrid model of an image processing technique with CNN that produced high resolution images into the features that were of the utmost importance in classifying the tumors. [7] and [8] focused on the importance of Region of Interest rather focusing on the entire tumor cells and had go computational speed as compared to other models. [9] imposed a novel method of genetic programming and machine learning algorithms which almost accurately defines the correct results.

3. MACHINE LEARNING MODELS

Machine Learning is the area of study that helps in computers the power or the capability of learning without actually being explicitly programmed and being a subset of AI, it is one of the most exciting technologies that the information domain has ever come across. It gives the computer the thought process similar to that of humans and that what makes it more similar to them. Machine learning is currently in active use, perhaps in many more places than one would expect. [1] and [3] suggest some good machine learning techniques which might be good from the optimization point of view.

ML techniques come across three categories which clearly specify them based on predicting outputs with certain input given.

1) Supervised learning: It is based upon the fact that whenever the input observations are fed into the algorithm in form a training data it generates a function that is clearly able to predict the output based on the testing data that is given to the system to predict.

2) Unsupervised learning: The machine is not given an input observation beforehand and therefore it is forced to learn from an unlabelled set of dataset to predict the output.

3) Reinforcement learning: The learning process in this category iterates over time and all the system states eventually learn input observations over a period of time.

3.1 Benchmark model

A Naïve Classifier model is used as a benchmark against our baseline models so as to provide a metric on the basis of which the comparison of performances of the sophisticated machine learning models are done. To get a measure of how well the models are doing for both of these at the same time. F1 score metric is used. Calculation of accuracy score is also done, although the F1 score will be the most important measure. The F1 score is also a good measure to use for relatively small dataset. A Naive Classifier is a simple classification model that simply assumes nothing about the underlying problem and the performance could be used as a baseline for complex algorithms through which all the other metrics of the dataset can be compared. There are different strategies that can be used for a naive classifier, and some are better than others, depending on the dataset and the choice of performance measures. The common performance of measure is classification accuracy and common naive classification strategies that includes randomly choosing labels.

3.2 Unoptimized baseline models

The following features were judged and found appropriate & desirable in guiding the selection of a machine learning model.

- Memory & speed**:** not primarily important but for what is feasible to be run within a notebook in a reasonable time, the aim is assumed to be a one off run of the model to predict rather than a model run continuously where significant computational resources might be more an issue.
- Not overfitting**:** A model that generalises well, its going to be run most likely periodically if used and to maximise the predictive power each time its run.
- Time for learning/fitting**:** like memory & speed, not so important beyond that which is reasonable to be run in the notebook without the browser timing out!
- Time for predicting**:** Its run on bigger datasets of patient dataset is a factor worth taking into account.

The above factors into account the following algorithms were considered as the baseline models:

Algorithm 1 - LogisticRegression

- Strengths of model: This algorithm i**s** fast to train with no parameter tuning and features don't need scaling, more tolerant to correlated features, excellent for 2-class classification problems.

- Why model is a good candidate for this problem**:** The problem is a binary classification one and logistic regression proves to be one of the best & simplest models for this kind of problems. The computational resources required for this model are low which proves to be a major advantage.

Algorithm 2 - GaussianNB (Gaussian Naive Bayes**)**

- Strengths of model**:** This algorithm i**s** easily implementable and has a high computational speedup. Also, a Naive Bayes classifier has a much higher performance than classifiers like logistic regression and because of the fact that noises are minimal to low and the training dataset required is of very small in size.
- Why model is a good candidate for this problem: This model could be useful if prediction on larger datasets for more patients is done, as it is likely to scale well. It is Good to evaluate another 'fast' model in addition to logistic regression.

Algorithm 3 - GradientBoostingClassifier

- Strengths of model**:** This algorithm can handle big datasets, very accurately and can approximate most non-linear classification boundaries, it is one of best boosting models for many classification problems (best in class)
- Why model is a good candidate for this problem**:** This model is one of the most popular (getting some of the best results) models for classification problems and also paramaterization can make a real difference to performance with this model with a good chance of improvement in optimisation/grid search phase for one of the best solutions (assuming computational resources of notebook/browser sufficient).

3.3 Optimized models

The best performing unoptimised model which is is tuned on various hyperparameters with F1-score as the scorer using grid search technique. Grid-searching is the process of searching those parameters which can give us best results whenever the tuning process is done. Depending on the type of model used some parameters are absolutely necessary such as human based factors. It not only applies to one model type but can be applied across machine learning to calculate the best parameters to use for any given model.

4. PROPOSED METHODOLOGY

4.1 Dataset Description

The dataset of patient data relating to breast cancer is available on Kaggle as the Wisconsin Breast Cancer dataset. A trained specialist can then decide if there is cancer or not. It consists of 32 features which relate to these types of images & cells and are the following:

- ID number
- Diagnosis of the tumour either benign or malign

The feature computed for each nucleus was:

- Mean of the distance of the centre with the outline.
- Deviation in values of gray-scale
- The total perimeter of the tumour
- The total area of the tumour
- Variance of the smoothness in tumours
- Formula derived from the area and perimeter
- Value defining the severe concave points
- The total number of concave points
- symmetry
- fractal dimension

There were no missing value attributes. The total no of class distribution contained 357 benign, 212 malignant data.

4.2 Proposed solution

Step 1: The Wisconsin dataset was first used for an initial data exploration technique where all the missing values data were removed and any kind of noises present in the dataset was replaced with mean values of the column. The thirty second column was redundant and hence was dropped for further use.

Step 2: The dataset was used for further exploration where a univariate exploration of each of the numerical features was done, by dividing the dataset into sub-groups for 'diagnosis' i.e. malignant and benign sub groups, with comparison of the distributions for each sub groups to see what patterns there are. These might prove useful features for our model to help better predict & distinguish between and predict malignant v benign tumors.

Step 3: From previous step, some distributions were found that were particularly skewed in nature. So the distributions were log-transformed to spread out the values more to make them of maximum use to the model. Thirdly, the units of the features and scale vary widely, with many such as 'symmetry_mean' being fractional decimals below 1, and others such as 'area_worst' has values ranging from a few hundred to over 3500. ML models are very sensitive to these differences of scale so to ensure the model treats all features equally, normalization of each feature was done, so that each feature is a value between 0 and 1.

Step 4: A different approach to get a perspective on the data and the differences between the benign and malignant groups was done using Topological Data Analysis (TDA). TDA is not a dimensionality reduction technique as such but rather, it produces a more abstract representation of the data (a 'simplicial complex' summary of the original data - an amusing mathematical oxymoron!). This complex could be said to represent the **'shape'** of the data, or a higher order

representation of the data. The topological network of the data reveals some interesting features.

Step 5: The normalized features of the dataset were then used in predicting the outcome of benchmark model which was a Naïve classifier. The performance from the benchmark model served as a metric for comparison.

Step 6: The best performing model from the three baseline models namely Logistic Regression, GaussianNB and Random Boosting classifier was then used to form our candidate model which could be then used for optimization.

Step7: The candidate model was optimized on hyperparameters through grid search methods and the results were them compared with the candidate models to compare its performance.

Step 8: The optimized model was then used to predict the outcome of the input data given along with it giving the five topmost features that were of utmost importance in predicting the label.

4.3 Metrics Used

1. Accuracy is used as a predictor that defines how accurate or how correct the values in comparison to the actual results defined. The equation to define it is presented below:

$$\frac{(\text{TruePositive} + \text{TrueNegetive})}{(\text{TruePositive} + \text{FalsePositive} + \text{TrueNegative} + \text{FalseNegative})}$$

2. F1 Score which is also known as the mean of Precision and Recall.F1 score is considered perfect and at best performing condition when at 1 and is a total failure when at 0.

$$\text{F1 Score} = \frac{2*(\text{Precision}*\text{Recall})}{(\text{Precision} + \text{Recall})}$$

5. RESULTS AND DISCUSSIONS

5.1. Comparison among proposed models

The observations for results of the model training and evaluation in the below plots the metric scores for each baseline model type. Firstly, all 3 models have done quite well, with all models getting at least 0.80 on the testing dataset, even when only 10% or 50% of that testing set was used. For the model training, the Gradient Boosting Classifier took by far the most time to train, and in terms of speed of predicting unsurprisingly Logistic Regression was the fastest. However, in terms of the key metric using 100% of the test dataset, the Gradient Boosting Classifier seems to perform the best. Therefore, we will select the Gradient Boosting Classifier as our best model and the one we want to optimize going forward.

Figure-2: Comparison of baseline models

Table-1: Summary table of the key results from developing the ML model.

	ACCURACY	F1-SCORE
BENCHMARK MODEL	0.3726	0.4280
CANDIDATE MODEL	0.9649	0.9574
OPTIMIZED MODEL	0.9912	0.9957

The candidate model performs significantly better on both accuracy and F1-score than the benchmark model. Also noticing is that the further parameter tuning with grid search made even more gains on both metrics to produce an **exceptional F-score of 0.9957** for the optimized model.

5.2. Topological Data Analysis of the labels present

The plot below shows the topological network of the dataset, with the malignant group colored in yellow, the benign group colored in purple, and the other colors various mixtures of malignant and benign cases. The plot is actually a fully animated and interactive network that can be explored using mouse to view the full interactive version of the network, this network is a more abstract and higher level summary of all the features combined, with the malignant and benign groups colored. First point noted from the visualization is that the benign group is more tightly packed, indicating there is less variance in the values of the features for this group. Second, and conversely, the observation

specifies that the malignant group is far more diverse, including many levels of distinct sub-sub groups within it. This indicates that there is a higher variance in the values of the features for this group. There is also a mixed group where the two main groups overlap, but this is relatively small.

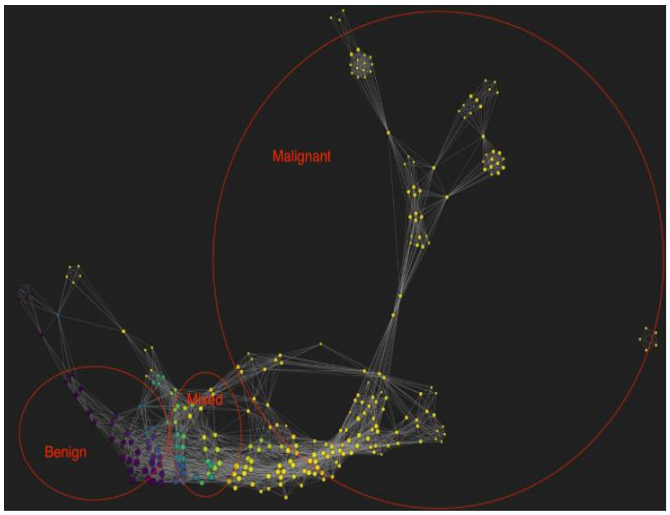

Figure-3: K-Mapper visualisation of features present in the two labels

5.3. Extracting the topmost features from the final model

The extracted top 5 features used by the final model were:

- concave points_worst
- perimeter worst
- concave points_mean
- area_worst
- texture_worst

Interestingly, these were features were highlighted earlier in the univariate analysis as features that had a very different distribution of values for the malignant and benign tumour sub-groups.

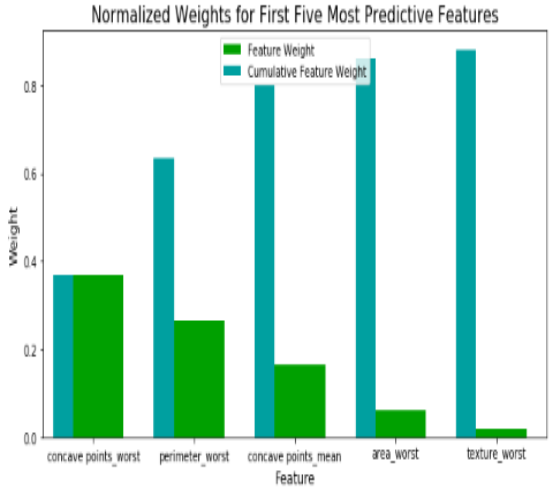

Figure-4: Normalized weight plotting of the extracted topmost features

6. CONCLUSIONS

In this paper, there has been close observation towards the Wisconsin dataset relating to breast cancer, and a fully developed model that is able to predict malignant tumors with a very high degree of accuracy (an F-score of 0.9957). There was a keen understanding of the reasons the proposed model is able to predict this well. The earlier analysis showed the difference in morphology between cell metrics for malignant v benign tumors, which could be seen visually in the images and were expressed in different distributions of values for particular feature measurements of the cells that we observed.

Use of higher level analytical tools such as TDA also allowed to gain a much better understanding of the dataset, in particular a better idea of the range of feature values that were typical for malignant and benign tumors as 'groups' within the dataset.

REFERENCES

[1] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury. "Breast Cancer Detection Using Machine Learning Algorithms", 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018.

[2] Minghao Piao. "Discovery of Significant Classification Rules from Incrementally Inducted Decision Tree Ensemble for Diagnosis of Disease", Lecture Notes in Computer Science, 2009.

[3] T Choudhury, V Kumar, D Nigam, B Mandal ,Intelligent classification of lung & oral cancer through diverse data mining algorithms, International Conference on Micro-Electronics and Telecommunication Engineering 2016.

[4] "CSITSS Proceedings 2020", 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2019.

[5] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", *Biol Cybern*, vol. 36, no. 4, pp. 193-202, 1980.

[6] M. Negnevitsky, Artificial Intelligence A Guide to Intelligent Systems, England:Pearson Education Limited, 2002.

[7] https://breast-cancer-research.biomedcentral.com/articles.

[8] https://www.ncbi.nlm.nih.gov/.

[9] Seker H, Odetayo MO, Petrovic D, et al. 2002. Assessment of nodal involvement and survival analysis in breast cancer patients using image cytometric data: statistical, neural network and fuzzy approaches. Anticancer Res, 22:433-8.