A Multi-syndrome Pathology for Breast Cancer through Intelligent Learning

Dr. C.N. Vanitha¹ and S. Malathy²*
¹,²Assistant Professor, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai - 638 060
Email: ksmalathy@gmail.com

Abstract: The field of data analytics dealt with various methods to analyze the available data and help us to draw into a point of conclusion about that information. Machine will learn the behavior based on the algorithms applied on those data and provide an appropriate conclusion. When these kinds of data analytic metrics applied to medical diagnosis various symptoms obtained from the patients and then the disease could be detected effectively. When certain optimization methods are applied on the data, the overall efficiency of the system will be improved considerably. To meet the challenges in diagnosing breast cancer based on clinical record sources, different symptoms in descriptions, clinical symptoms, a novel method, which consists of choosing the suitable features, multi-class functions, and multi-label parameters, has been proposed. The proposed work will be implemented as two steps such as discriminative symptoms selection and multi-syndrome learning. Public Breast Cancer Wisconsin (Diagnostic) Data Set has been taken for implementation. The breast cancer data set utilized for this work comprises 699 tumour samples. In that 458 (65.5%) samples belong to Benevolent (non-cancer) tumours and 241 (34.5%) belongs to Malevolent (cancer) tumours. The overall verisimilitude is 95.23% which will be improved to a greater extent when compared to the existing schemes.

Keywords – Breast cancer, clinical symptoms, techniques, machine learning, accuracy

1. Introduction

Machine learning provides smarter solutions to health related applications. Artificial intelligence (AI) helps the machine to learn the behavior of the user by using voice assistants and analyzing the online shopping experience by implementing learning association algorithms. Integrating machine-learning techniques especially with healthcare applications provides better solutions in medical diagnosis. Now-a-days machine learning is growing faster in healthcare. But there is a biggest challenge in making the patient data set to fit into the suitable size and quality. Those samples will be trained to the novel and effective machine learning models [1]. The in-patient and out-patient data has been protected strictly with security procedures so that the data are not easy to collect, share and distribute.

In order to attain the desired format and quality of the data, more effort is required to prepare the analysis from data available in internet using machine learning which has to consider its elements and challenges. The machine learning slowly adapts the tools in the healthcare applications and the industries are pushing to those boundaries. Its primary function is to analysis the data of each patient in terms of X-rays, genetic Deoxyribonucleic Acid (DNA) sequences, descriptions about vaccinations, blood test reports, current medicine and other past medical histories. By implementing classification techniques in general applications for a large scale data could improve the predictive verisimilitude [16]. It plays significant role in many real time applications especially in healthcare applications as follows.

Chatbots:

AI chat bots are mainly used in companies by incorporating speech capability to find the patient’s symptoms and generate potential diagnosis to prevent disease and/or recommend an appropriate treatment to be taken [19].

Oncology:

Now-a-days scientists are using internet of things [17] and deep learning techniques to design an appropriate algorithm for detecting the level of cancer-infected tissue. It is more useful for the physician to compare the reports.

Pathology:

Pathology is the specialty in the medicine that is worried with the disease diagnosis based on the processing in the laboratory. It analyses the bodily fluids such as Red Blood Corpuscles (RBC), White Blood Corpuscles (WBC), platelets, urine and some tissues. The machine vision and the machine learning schemes has been integrated with the traditional efforts, which improves the results far better.
2. Literature review

The technique to analyze face based on both permanent and transient facial expressions has been developed. Instead of macro prototyping parts fine-grained parts called action units were considered. Tracking has been done either by individual or a group of action units considered [1]. Evolutionary optimization technique has been used to reduce the space, cost and dimension of the features considered and verisimilitude of the results obtained are higher than the traditional methods [8]. Features are arranged in a number of clusters and a decision rules has been constructed for each cluster. The final decision about the disease could be arrived based on the rules applied [9].

Statistical methods are used for grouping various kinds of data such as sound, color and words. The size of the image should be equal in training and testing dataset to keep the features constant [13]. A special kind of Support Vector Machine (SVM) using least square method helps to detect breast cancer. Verisimilitude has been achieved more using cross-validating the matrix [14]. Grid search method is combined with f-score to give the predictive and classification verisimilitude [15].

A gray level co-occurrence matrix and neighbor region dependent feature has been utilized to get the results progressive [18]. Self Adaptive Network has been designed to detect the abnormal growth in breast using digitized mammograms. The distortion of muscles leads to breast cancer has been identified successfully [19].

3. Statistical report

There is a huge rise in the number of breast cancer cases in our India. There is no way to control the fear of breast cancer but can control the numbers by treating them at the right time (at the initial stage). Breast Cancer remains 25% to 35% in all the cities all over India. For every 13 minutes, a female dies for breast cancer in India. According to the statistical report, 70,200 women died only because of breast cancer at the year 2012. This huge rate is only because of the lack of awareness and delay in diagnosing the disease. More than 45% of women suffering in India from third and fourth stage so that their chance of survival rate is very low. Figure 1 indicates that the age group belongs to 50-60 are suffering a lot from this type of cancer. With the help of the study reports, it is known that Kerala is having the highest breast count rate around India. Haryana plays the second role. Women in the early thirties to fifties are at the risk in developing breast cancer. 2018 records show that 1,62,000 new cases were reported and 87,000 died due to that cancer. Self-diagnosing can be done according to their condition by their own. It is possible to increase the survival rate of the people who are infected with breast cancer when the disease is diagnosed in the correct time. It can be controlled by self-examining their breasts accordingly after the age of 30. Lack of doing physical exercises, consuming junk foods leads to obesity which in turn leads to the breast cancer. According to the report published by United Nations Organisation (UNO) states that there is a massive increase in the growth of breast cancer this year.

![Figure 1. Statistical Report on the growth of Breast Cancer cases](image)

4. Incorporating machine learning for diagnosing breast cancer
Main applications of machine learning in health care are the digital diagnosis [2]. Machine Learning detects the vulnerable patterns automatically within the patient’s electronic health care records and they inform to the clinicians in case of emergency. Artificial intelligence can also evaluate the patient’s health record depending on the knowledge acquired from wider data sets. It is capable of analyzing the millions of disease that the patient can have. According to this, we can use Machine Learning (ML) for the purpose of detecting the breast cancer[8] and that is available in Public Breast Cancer Wisconsin (Diagnostic) Data Set.

Figure 2. Breast cancer tissues

This data set contains the various objects of the tumor. Tumours are two types; they can be either a benevolent or malevolent. Benevolent tumour can grow locally, Malevolent tumour spreads and infect other healthy tissues. It causes severe damage in the vital organs like brain[6]. This process of affecting other tissues by spreading is called as metastasis There are various types of tumours with malevolent nature are available and their locations where they are originated are described as data set specification.

5. Data set on breast cancer

The features considered in the dataset are sample code number which is unique in nature. The thickness of the clump, cell size, single epithelial cell size, margin or the boundary of adhesion, location of the nucleus of the cell, single epithelial cell, chromatin, size of the normal nuclei are considered for computation. The dataset is splitted into two important parts for training and testing purpose. Metrics have been defined for positive, negative, wrong positive and wrong negative cases. The values of verisimilitude, verisimilitude, revocation has been calculated. The models are then evaluated and an neural network has been designed.

Figure 3. Attribute and domain values

The breast cancer data set utilized for this work comprises 699 tumour samples. In that 458(65.5 %) samples belong to Benevolent (non-cancer) tumours and 241(34.5%) belongs to malevolent(cancer) tumours[3] as shown in Figure 2. To solve this, we have to use some machine learning methodologies to detect the breast cancer, the following are the set of parameters are used: ROC curve ~0.99[4], verisimilitude— Revocation curve ~ 0.99 and F1~0.97

6. A novel diagnosis model

A novel diagnosis model has been proposed as shown in Figure 4. The proposed model consists of the following steps.

**STEP 01:** Dividing the Data Set

1. Train set (i.e. 65% of the repository) is used for selecting and validating the models, and
2. Test set (i.e. 35% of the repository), hold the data and we can see how well the models are generalizing the unseen data [5].

We can define the key metrics to measuring the efficiency of the models [7]. We have the most important four characteristics to define the derivative measurements metrics.

**TEP 02: Defining the metrics**

The four characteristics are: Let us assume the meaning of

α - Right and β - Wrong.

1. αP (Right Positive) – It refers to the actual count of the patients who are correctly categorized as having the diseases,
2. αN (Right Negative) – It refers to the actual count of the patients who are correctly categorized as having good health condition,
3. βP (Wrong Positive) – It counts the number of patients who are actually healthy but wrongly classified as cancer infected patients.
4. βN (Wrong Negative) – It counts the number of patients who are wrongly identified to have the disease.

**Figure 4. Flowchart of diagnosis model**

Depending upon these specifications the following parameters has been defined:

1. **Verisimilitude:**
   
   It is the ratio of sum of right positive patients and right negative patients to the total number of patients considered [9].

   \[
   \text{Verisimilitude} = \frac{\sum (\alpha_p + \alpha_n + \beta_p + \beta_n)}{\sum \text{(Number of patients)}}
   \]  
   \[(1)\]

2. **Exactitude:**

   It is the ratio of sum of right positive and wrong negative patients with the disease to the total patients classified as they are having the disease. The Strategy behind verisimilitude is to find out how many patients classified as having disease truly.

   \[
   \text{Exactitude} = \frac{\sum (\alpha_p + \beta_n)}{\alpha_p}
   \]  
   \[(2)\]

3. **Revocation:**

   It is the ratio of patients who were classified as right positives to the patients who have the diseases. The overall aim behind this revocation is how many patients who have disease classified as having the disease [19].

   \[
   \text{Revocation} = \frac{\sum (\alpha_p + \beta_n)}{\alpha_p}
   \]  
   \[(3)\]

**STEP 03: Evaluating the models**
The F1 score is estimated as the ratio between twice the product of verisimilitude values and revocationing values to the sum of verisimilitude and revocation values. These are the essential metrics which are used to evaluate the models [10].

\[
(F1 \text{ score} = \frac{2 \times (\text{revocation} \times \text{exactitude})}{\text{revocation} + \text{exactitude}})
\]  

(4)

The cross confirmation value of the ensemble ML models will provide the promising decision boundaries as output.

**STEP 4: Neural Network Model Creation**

![Figure 5. Cross Verification result of top 3 models](image)

This model gives the value of about .97 F1 on average by cross verification. The value of F1 in this model (Neural Network)[11] is better when compared with value of the best model obtained in the step 3. The top results of the model are cross verified with three models such as Gaussian process, gaussianNB_esbl, NN model. Lets now evaluate those models on the bases of test data which not shown previously for the classifiers for imitating the new data. The below diagram shows how better the models are performed with the help of test data [15]. According the graph, the classifiers of neural networks have been worked well with the gain of .96 F1 value on the test dataset [12].

**STEP 5: Receiver Operating Curves for Evaluating Output Quality:**

(ROC) Receiver operating characteristic curves is used for further evaluation of classifier’s output quality [16]. The Receiver Operating Characteristic efficiency graph has been evaluated by the surface area below the curve. The value .1 under the area curve represents the classifier is perfect and the value .5 shows the valueless classifier. In table 1, the range, classifier and its label is defined. The scale ranges from 0 to 1. There are five different classifiers named as (A, B, C, D, F) available in that range. If the value is between 1 to 0.9 it is classified as perfect. If the value is less than 0.5 it is noted as not succeeded.

![Figure 6. ROC Characteristics](image)

**Table 1: Classifier table**
### Table 1: Classification Scores and Labels

| Range     | Classifier | Label |
|-----------|------------|-------|
| 0.9-1     | Perfect    | A     |
| 0.80-0.90 | Fine       | B     |
| 0.70-0.80 | Light      | C     |
| 0.60-0.70 | Pitiable   | D     |
| 0.50-0.60 | Not succeed| F     |

The aforementioned are the academic points for judging the effectiveness of the classifier. The graph shows the top three classifiers having .99 areas under the area which is consider as perfect.

**STEP 6: Exactitude Revocation Curves for Evaluating the Quality of Output**

The navy colored dash on the graph represents 1 average exactitude and seeing the top 3 models the average exactitude is almost equal to 1, which is the perfect value[13].

![Figure 7. Exactitude revocation curves for top three models](image)

**STEP 7: Visualization of Decision Boundaries:**

The last view of the model is decision boundary [14] To attain more visual perception about the test data and the algorithms, this decision boundary reduces the 9D dimensionality spaces to 2D with the help of visualize boundary and Principal Component Analysis (PCA) technique.

![Figure 8. Visualizing decision boundaries](image)

### 7. Conclusion
The integration of data analytics and machine learning in health care applications plays the most significant role in today’s world. The proposed novel method consists of discriminative symptoms selection and multi-syndrome learning which helps us to detect the cancer in an efficient way. The proposed novel diagnosis model produces the perfect results with the verisimilitude rate of 95.23% on breast cancer detection in the data set considered. Though the model gives the expecting results but it has to be further strengthened for the clinical use by including genetic sequence, demography and race to make the ground stronger for classification.

8. REFERENCES

[1] Tian TK, Cohn JF 2001 Recognizing action units for facial expression analysis IEEE Transactions on Pattern Analysis and Machine Intelligence 23 97-115 doi: 10.1109/34.908962
[2] Sen SK 2019.
[3] Christopher MB 2010.
[4] Yaser S, Abu M 2012.
[5] Mukharjee KL 1997.
[6] DSM-5® Diagnostic and Statistical Manual of Mental Disorders publication 2013.
[7] Maryam V, Vahid A, Reza F, 2013 Perceptual pain classification using ANFIS adapted RBF kernel support vector machine for therapeutic usage Applied Soft Computing 13 2537-2546 doi:10.1016/j.asoc.2012.11.032
[8] Fan CY, Chang PC, Lin JJ, Hsieh JC, 2011 A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification Appl. Soft Comput 11 632–644.
[9] Escalante HH, Montes-y-Gómez M, González JA, Gómez-Gil P, Altamirano L, Reyes CA, Reta C, Rosales A, 2012 Artif. Intell. Med., 55 163–175.
[10] Sana S et al, 2019 Assessment of Machine Learning Detection of Environmental Enteropathy and Celiac Disease in Children, JAMA Network Open.
[11] Brink H, Richards JW 2013 Real-World Machine Learning 7th ed Manning Publications.
[12] Deepa SN, Devi BA 2011 A Survey on Artificial Intelligence Approaches for Medical Image Classification Indian J. Sci. Technol 4 1583–1595.
[13] Polat K, Güneş S 2007 Breast cancer diagnosis using least square support vector machine Digiti. Signal Process., 17 694–701.
[14] Chen HL, Yang B, Wang G, Wang SJ, Liu J, Liu DY 2011 Support Vector Machine Based Diagnostic System for Breast Cancer Using Swarm Intelligence J. Med. Syst 36 2505–2519.
[15] Othman DMF, Yau TMS 2006 Comparison of Different Classification Techniques Using WEKA for Breast Cancer in 3rd International Conference on Biomedical Engineering Kuala Lumpur (Springer) pp 511.
[16] Ibrahim F, Osman NAA, Usman J, Kadri NA 2007 Springer Berlin Heidelberg 520–523.
[17] Dhiviya S, Malathy S, Rajesh Kumar D 2018 Internet of Things (IoT) Elements, Trends and Applications, International Journal on Computational and Theoretical Nanoscience 15 1639 - 1643.
[18] Shanthi, Murali B 2013 A Novel Approach for Detecting and Classifying Breast Cancer in Mammogram Images International Journal of Intelligent Information Technologies 9 21-39.
[19] Shanthi, Murali B 2014 A novel approach for classification of abnormalities in digitized mammograms Sadhana Academy Proceedings in Engineering Science journal 39 1141–1150.