Reliable Probabilistic Classification of Mammographic Masses using Random Forests

Hechmi Shili[1,2], **Lotfi Ben Romdhane**[1,3], **and Béchir el Ayeb**[2]
[1]MARS Research Group, Faculty of Sciences of Monastir
[2]University of Monastir, Monastir, Tunisia
[3]High School of Sciences and Technology, Hammam-Sousse, University of Sousse

Abstract— *Mammography is the most effective method for identifying breast cancer in its earliest stages. Random forests (RF) have been successfully used for the task of classification with good performance, but without information about the reliability in classifications. In this paper, we present a novel reliable probabilistic approach to classify mammographic masses as benign, malignant and normal tissues. The main aim of this paper is to improve the performance of Random forests by introducing a recently developed algorithmic framework, namely the Venn Probability Machine, for making reliable decisions in the face of uncertainty.*

Keywords: Mammography, Probabilistic classification; Random forests; Venn prediction.

1. Introduction and Background

Breast cancer is the most common cause of cancer-related death in women worldwide, with some 327 000 deaths each year. Nearly 1.4 million cases of breast cancer were diagnosed across the world in 2008, compared with about 500 000 cases in 1975. This represents about 11% of all new cancer cases and 23% of all female cancers. It is predicted that the number of cases will rise to 1.7 million by 2020 [6]. Primary prevention seems impossible since the causes of this disease are still remaining unidentified. Early detection is the key to the ultimate survival rate for breast cancer patients. For women whose tumors were discovered early, the five year survival rate was about 82%, as opposed 60% that not been found early [6].

Mammography is still the most effective screening method for detecting breast cancer in its early, most treatable stages. However, the low positive predictive value of breast biopsy examinations resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies performed on benign lesions. Computer-aided diagnosis (CADx) systems have been developed to assist the radiologist in the discrimination of benign and malignant breast lesions and thus to reduce the high number of unnecessary biopsies. It is important to realize that the classification of suspicious abnormalities in digital mammograms is an extremely challenging task for a number of reasons. First, it is a challenge to select a good feature set for the classification of mammogram. Second, abnormalities are often occluded or hidden in dense breast tissue, which makes detection difficult. Finally, symptoms of abnormal tissue may remain quite subtle. For example, speculated masses that may indicate a malignant tissue within the breast are often difficult to correctly diagnose, especially at the early stage of development.

As such, an increasing number of researchers have focused on the classification of suspicious masses in mammograms. Quite a lot of researchers apply the classification techniques to classify the marked region in the mammogram. Classifiers like decision tree classifiers [15], [8], Support Vector Machines [9], [16], k-nearest neighbors [12] and Artificial Neural Network [7], [1] have performed better in mass classification.

Most of the methods mentioned previously provide too little insight as to the importance of variables to the predictor derived. The transparency is very important in some application areas such as medical decision support. By contrast, classification and regression trees are known for their transparency. Decision tree have been widely and successfully used in mammographic mass classification.

In [15], the authors used a method based on binary trees for the classification of mammograms. Global feature extraction from different levels wavelet decomposition of normal and abnormal images was also used in this work. This classifier is then used to classify whether an entire whole-field mammogram is normal. However, in such a binary tree classifier, errors may accumulate from one level to another, thus making the classification erroneous. Hence, this method resulted in false positive in more than 50% of the cases, making it unreliable.

In [8], the authors discusses the effectiveness of using decision trees for mass classification in mammography. Different costs for type I and type II misclassification were applied for the experiments. The results obtained using algorithms based on decision trees were compared with that produced by neural network which was reported giving the higher classification rate than statistical models, with higher standard deviation. It is concluded that the decision trees are very promising for the classification of breast masses in digital mammograms. However, decision trees

are rather unstable: small changes in the training set can result in different trees and different predictions for the same validation examples. It has been demonstrated that this problem can be mitigated by applying bagging [4]. Random Forests (RF) proposed by Breiman [4] is a combination of the random subspace method and bagging.

In [17], an approach using Random Forests Decision Classifier (RFDC), involving regression trees, has been used in mammogram classification. The technique in [17] yielded an accuracy of nearly 90%. However, this method is not very reliable as features are randomly selected in the tree induction process.

In [10], the authors investigated the usage of Random forests classifier for the classification of masses with geometry and texture features. The experiments are tested using a database of 236 clinical mammograms. This method achieved an average area under the ROC curve of 0.86 with Support Vector Machines (SVM) and 0.83 with Random forests. The experimental result shows that Random forests is a promising method for the diagnosis of masses.

Meyer et al. [11] compared 17 classifiers on 21 datasets obtained from the above-mentioned repository. RF outperformed neural network in terms of average test set errors in 15 cases, SVM in 7 cases. The authors concluded that ensemble methods - such as RF - proved very competitive, and often produce adequate results "out of the box", whereas SVM react very delicately to parameter tuning.

However, like most machine learning algorithms, Random forests outputs the label predictions for new instances without indicating how reliable the predictions are. The applicability of these classifiers is limited in critical domains where incorrect predictions have serious consequences, like medical diagnosis. Further, the default assumption of equal misclassification costs is most likely violated in medical diagnosis. This paper addresses the importance of reliability and confidence for classification, and presents a novel method based on a combination of Random forests, and Venn Prediction (VP) [18].

Venn Prediction is an extension of the original conformal predictor (CP) framework , which can be used for making multiprobability predictions [18]. In particular multiprobability predictions are a set of probability distributions for the true classification of the new example. This set can be summarized by lower and upper bounds for the conditional probability of the new example belonging to each one of the possible classes. The resulting bounds are guaranteed to contain well-calibrated probabilities (up to statistical fluctuations). Again, like with CPs, the only assumption made for obtaining this guaranty is that the data are are generated independently by the same probability distribution (i.i.d). The VP framework has until now been combined with the k-nearest neighbours algorithm in [18], [5], with SVMs in [19] and more recently with Neural Networks in [13].

This work is aimed at improving performance of the current mass classification methods using Random Froest classifiers. The novelty of this research is in exploiting the superiority of Venn prediction to produce probability estimates that are guaranteed to be well calibrated. The rest of this paper is organized as follows: Section 2 describes about the Random forests method. Section 3, details the Venn Prediction framework. Section 4 presents our proposed Algorithm for classifying masses in breast. Section 5 describes the experiments that have been conducted on benchmark data set. Finally, Section 6 presents some concluding remarks.

2. Random Forests

Random Forests (RF) is an ensemble learning technique developed by Breiman [4]. This technique combines many decision trees to make a prediction, giving as output the class that is the mode of the classes output by individual trees.

RFs is a family of methods, made of different decision tree ensemble induction algorithms, such as the Breiman Forest-RI method often cited as the reference algorithm in the literature [4]. In this algorithm, the training set for each individual tree in a Random forests is constructed by sampling N examples at random with replacement from the N available examples in the dataset. This is known as bootstrap sampling, and bagging describes the aggregation of predictions from the resulting collection of trees. As a result of the bootstrap sampling procedure, approximately one third of the available N examples are not present in the training set of each tree. The "out-of-bag" predictions are those predictions derived from non-bootstrapped observations which built that particular tree.

In this induction algorithm, a feature subset is randomly drawn for each node, from which the best splitting criterion is then selected according to the Gini index (Breiman et al., [3]), which measures the likelihood that an example would be incorrectly labelled if it were randomly classified according to the distribution of labels within the node. For a binary split, the Gini index of a node n may be expressed as $I_G(n) = 1 - \sum_{c=1}^{2} p_c^2$, where p_c is the relative proportion of examples belonging to class c present in node n. Thus, the Forest-RI Algorithm grows a decision tree using the following process :

Let T be the number of trees to build, for each of $|T|$ iterations

1) Select a new bootstrap sample from training set.
2) Grow an un-pruned tree on this bootstrap.
3) At each internal node, randomly select try m predictors and determine the best split using only these predictors.
4) Output overall prediction as the majority vote from all individually trained trees.

Figure 1 illustrates the workflow for random forests, where y_1, y_2, ...; y_c are class labels. As more trees are added to RF, the generalization error converges to a limiting value, thus there is no over-fitting in large RFs [4].

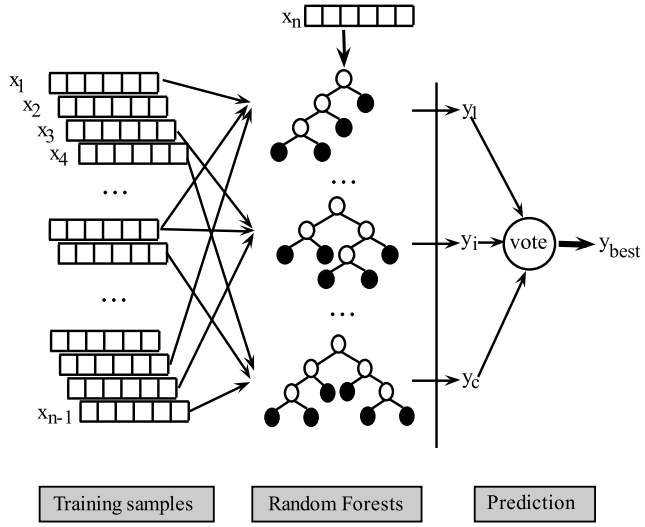


Fig. 1: General architecture of RF classifier.

The main advantage of Random Forests over other techniques such as Artificial Neural Networks, Support Vector Machines, Linear Discriminant Analysis, etc. is the robustness of this technique regarding solution over fitting, tending to converge always when the number of trees is large.

To assess the importance of a specific predictor variable (feature), the values of the variable in the out-of-bag samples are randomly permuted and then the modified out-of-bag samples are passed down the tree to get new predictions. The increase of estimation error for the modified and original out-of-bag data provides a useful measure for determining the feature importance, although feature selection is not needed in RF (Breiman and Cutler, [2]).

3. The framework of Venn machines

This section provides a brief overview of the Venn prediction mechanism; for more details the interested reader is referred to [18].

Let us consider a training set consisting of examples $Z = \{(x_i, y_i)\}_{i=1}^{n-1}$, where each $x_i \in \mathbb{R}^d$ is the vector of attributes for example i and $y_i \in Y = \{y_j\}_{j=1}^{c}$ is the class label of that example. Let x_n be a new unclassified example. Our task is to predict the probability of this new example belonging to each class $y_j \in Y$ based only on the assumption that all (x_i, y_i), $i = 1, 2, \cdots$ are generated independently by the same probability distribution (i.i.d).

The essential idea of Venn prediction is to divide all examples into a number of categories based on their similarity and calculate the probability of x_n belonging to each class $y_j \in Y$ as the frequency of y_j in the category that contains it. Then since we do not know the true labels of the new object x_n, we try every possible label as a candidate for its label. In each try, we calculate a probability distribution for the true class of x_n based on the examples

$$\{(x_1, y_1), \cdots, (x_{n-1}, y_{n-1}), (x_n, y_n)\}. \quad (1)$$

To divide each set (1) into categories we use a *taxonomy function*. $A_n : \mathbf{Z}^{n-1} \times \mathbf{Z} \to T, n \in \mathbf{N}$, which classifies the relation between an example and the bag of the other examples:

$$\begin{aligned}\tau_i = & A_n((x_i, y_i), \{(x_1, y_1), \cdots, (x_{i-1}, y_{i-1}), \\ & (x_{i+1}, y_{i+1}), \cdots, (x_n, y_n)\}).\end{aligned} \quad (2)$$

Values τ_i are called categories and are taken from a finite set $T = \{\tau_i, \tau_i, \cdots, \tau_k\}$. Equivalently, a taxonomy function assigns to each example (x_i, y_i) its category τ_i, or, in other words, groups all examples to a finite set of categories.

Typically each taxonomy is based on a traditional machine learning algorithm, called the *underlying algorithm* of the Venn predictor. The output of this algorithm for each attribute vector $x_i, i = 1, \cdots, n$ after being trained either on the whole set (1), or on the set resulting after removing the pair (x_i, y_i) (2), is used to assign (x_i, y_i) to one of a predefined set of categories. At this point it is important to emphasize the difference between the classes of the problem and the categories of a Venn taxonomy. These categories are assigned examples based on the output classification label of the underlying algorithm and not on the true class to which each example belongs. There fore the category corresponding to a given classification label y_j will contain the examples that the underlying algorithm "believes" to belong to class y_j, which are not necessarily the same as the examples that actually do belong to that class since the underlying algorithm might be wrong in some cases.

The conventional way of using Venn ideas was as follows. Categories are formed using only the training set. For each non-empty category τ, the empirical probabilities of an object within category τ to have a label y_j are found as

$$P_\tau(y_j) = \frac{N_\tau(y_j)}{N_\tau}. \quad (3)$$

Where N_τ is the total number of examples from the training set assigned to category τ, and $N_\tau(y_j)$ is the number of examples within category τ that are labelled with y_j.

Now, given a new object x_n with the unknown label y_n, one should assign it somehow to the most likely category of those already found using only the training set; let $\tau^\star$ denote it. Then the empirical probabilities $P_{\tau^\star}(y_j)$ are considered as probabilities of the object x_n to have a label y_j. The idea of confidence machines allows us to construct several probability distributions of a label y_j for a new object. First we consider a hypothesis that the label y_n of a new object x_n is equal to y $(y_n = y)$. Then we add the pair (x_n, y) to the training set and apply the taxonomy function A to this extended sequence $\{(x_1, y_1), \cdots, (x_{n-1}, y_{n-1}), (x_n, y)\}$. Let

$\tau^\star(x_n, y)$ be the category containing the pair (x_n, y). Now for this category we calculate, as previously, the values $N_{\tau^\star}$, $N_{\tau^\star}(y_j)$ and empirical probability distribution

$$P_{\tau^\star(x_n,y)}(y_j) = \frac{N_{\tau^\star}(y_j)}{N_{\tau^\star}}, y_j \in Y. \quad (4)$$

This distribution depends implicitly on the object x_n and its hypothetical label y. Trying all possible hypotheses of the label y_n being equal to y, we obtain a set of distributions $P_y(y_j) = P_{\tau^\star(x_n,y)}(y_j)$ for all possible labels y.

The taxonomy used is still very important as it determines how efficient, or informative, the resulting predictions are. We want the diameter of multiprobability predictions and therefore their uncertainty to be small, since saying that the probability of a given classification label for an example is between 0.8 and 0.9 is much more informative than saying that it is between 0 and 0.9. We also want the predictions to be as close as possible to zero or one, indicating that a classification label is highly unlikely or highly likely respectively.

The maximum and minimum probabilities obtained for each classification label y_j define the interval for the probability of the new example belonging to y_j:

$$\left[\min_{y\in Y} P_{\tau^\star(x_n,y)}(y_j), \max_{y\in Y} P_{\tau^\star(x_n,y)}(y_j)\right]. \quad (5)$$

To simplify notation the lower bound of this interval for a given class y_j will be denoted as $L(y_j)$ and the upper bound will be denoted as $U(y_j)$. The Venn predictor outputs the best class $\hat{y}$ for x_n where:

$$\hat{y} = \arg\max_{j=1,\cdots,c} \overline{P(y_j)}. \quad (6)$$

and $\overline{P(y_j)}$ is the mean of the probabilities obtained for y_j:

$$\overline{P(y_j)} = \frac{1}{|Y|}\sum_{y\in Y} P_{\tau^\star(x_n,y)}(y_j). \quad (7)$$

This prediction is accompanied by the interval:

$$[L(\hat{y}), U(\hat{y})]. \quad (8)$$

as the probability interval of it being correct. The complementary interval

$$[1 - L(\hat{y}), 1 - U(\hat{y})]. \quad (9)$$

gives the probability that $\hat{y}$ is not the true classification label of the new example and it is called the error probability interval.

4. The Algorithm

The difference between alternative Venn Prediction methods is the taxonomy they use to divide examples into categories. Here a RF classifier defined which allocate examples into categories. This section describes our algorithm for reliable probabilistic classification of mammograhic masses. The main idea of the proposed Algorithm is to embed random forests in confidence machines. In this way, we expect designed Venn machines to inherit advantages of random forests.

First, we train a RF classifier, according to Forest-RI Algorithm, on the extended set (1). Second, we assign (x_i, y_i) to the corresponding category τ_i according to RF outputs $\{o_i^1, \cdots, o_i^l\}$. The predicted class of (x_i, y_i) is calculated by its majority vote of the out-of-bag predictions. Algorithm 1 presents the complete $RPRF$ algorithm.

Algorithm 1: Reliable Probabilistic Random forests (RPRF)

Input: Training set $Z = \{(x_i, y_i)\}_{i=1}^{n-1}$ in wich
$x_i = \{x_i^1, \cdots, x_i^d\} \in \mathbf{R}^d$ and
$y_i \in Y = \{y_1, \cdots, y_c\}$ the possible class for x_i,
x_n a new example to be classified.

Result: The best class for x_n : $\hat{y} = \arg\max_{j=1,\cdots,c} \overline{P(y_j)}$,
the probability interval for $\hat{y}$:
$\left[\min_{y\in Y} P_{\tau^\star(x_n,y)}(\hat{y}), \max_{y\in Y} P_{\tau^\star(x_n,y)}(\hat{y})\right]$

begin
  for $k \leftarrow 1$ **to** c **do**
    Train a random forest (RF) classifier, according to Forest-RI Algorithm, on the extended set $\{(x_1, y_1), \cdots, (x_{n-1}, y_{n-1}), (x_n, y_k)\}$;
    Supply the input patterns $x_1, \cdots, x_n$ to the trained RF to obtain the output values $\{o_1, \cdots, o_n\}$ based on the out-of-bag predictions;
    for $i \leftarrow 0$ **to** n **do**
      According to RF outputs $\{o_i^1, \cdots, o_i^c\}$, assign (x_i, y_i) to the corresponding category τ_i.
    end
    Find the most likely category that contains (x_n, y_k), let $\tau^\star$ denote it.
    for $j \leftarrow 0$ **to** c **do**
      Compute the empirical probability $P_{\tau^\star(x_n,y_k)}(y_j)$ using equation (4).
    end
  end
  for $j \leftarrow 0$ **to** c **do**
    Compute the mean of the probability $\overline{P(y_j)}$ using equation (7).
  end
end

Applying a RF classifier that was trained on the whole

training data set (1), the examples are divided into categories for each assumed classification label $y_k \in \{y_1, \cdots, y_c\}$ of x_n and the process described in section 3 is followed for calculating the outputs of the Reliable Probabilistic Random Forests (RPRF). The predictions are based on the out-of-bag predictions from the RF.

In the next section, we will analyze experimentally our proposed model.

5. Experimentation

In this section, we will analyse experimentally our proposed model to well-known other proposals using a standard reference database. Experimental settings and results are described in the sequel.

5.1 Experimental settings

To evaluate our method, we used mammograms from the Mammographic Image Analysis Society (MIAS) database [14]. Films were taken from the United Kingdom National Breast Screening Program; digitized to 50 micron pixel edge, and presented each pixel with an 8-bit word. The MIAS database consists of totally 322 digital mammograms from 161 patients, which belong to three big categories: normal, benign and malign. There are 208 normal, 63 benign and 51 malign images. The normal ones are those characterizing a healthy patient, the benign ones represent mammograms showing a tumor, but that tumor is not formed by cancerous cells, and the malign ones are those mammograms taken from patients with cancerous tumors. This database provides for each mammogram a meta-data from radiologists about the characteristics of background tissue, the type and the severity of abnormality and the coordinates of center; etc. Using this informations, suspicious regions with the given centre and radius have been extracted as the Regions of Interest (ROIs).

We use a set which consists of totally 285 ROIs, which belong to three categories: normal, benign and malign. There are 130 normal, 75 benign and 80 malign ROIs. The images from the MIAS dataset are separated for training and testing. The training ratio is set as 80%, i.e. 80% of the samples for training and 20% for testing.

The computer classification results were validated using the following standard criteria: Accuracy (AC), Sensitivity (SE) or Recall, Specificity (SP), the area under the ROC curve (Az), F-measure (F1), Precision (Prec), Brier Score (BS) and Matthews's correlation coefficient (MCC). These measures are calculated from confusion matrix. The confusion matrix describes actual and predicted classes of the proposed method and shown in table 1. Calculations of those performance measures were carried out as follows:

$$SE = TPR = \frac{TP}{(TP + FN)} \tag{10}$$

$$SP = 1 - FPR = \frac{TN}{(TN + FP)} \tag{11}$$

$$AC = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{12}$$

$$Prec = \frac{TP}{(TP + FP)} \tag{13}$$

$$F1 = 2 \times \frac{(Prec \times SE)}{(Prec + SE)} \tag{14}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{((TP + FP)(TP + FN)(TP + FP)(TN + FN))}} \tag{15}$$

where FP, FN, TP and TN denote false-positive, false-negative, true-positive and true-negative answers, respectively. Moreover, FPR and TPR denote false-positive rate and true-positive rate, respectively.

The Brier score, BS, is defined for a dichotomous event as the mean square error of the probability forecast:

$$BS = \frac{1}{M} \sum_{i=1}^{M} (p_i - o_i)^2 \tag{16}$$

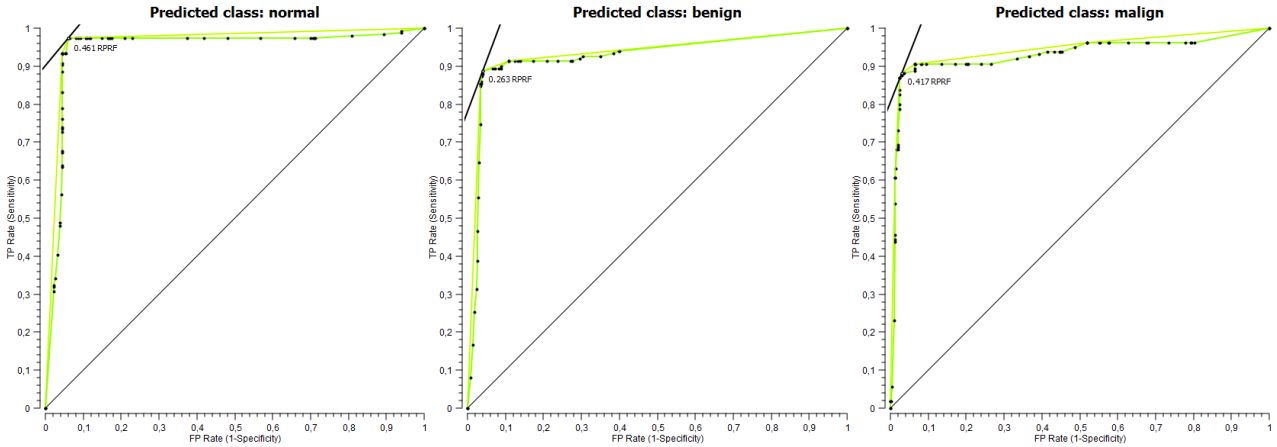
where M is the total number of samples, p_i is the forecast probability, o_i is the verifying observation (1 if the event occurs, 0 if it does not).

5.2 Comparative analysis

The classification performance of the proposed system is compared with that of other three existing classifiers like Support Vector Machine (SVM) [16], Probabilistic neural network (PNN) [1] and Random Forests (RF) [17] classifiers. Numerical results are summarized in Tables 1 and 2.

Table 1 shows the confusion matrices for all used classifiers. This should be read as follows: rows indicate the object to be recognized (the true class) and columns indicate the label the classifiers associates at this object, thus obtaining the correct classified mammograms in the diagonal of the matrix. Therefore, the performance of this approach is 91.92%. We can see that the mammograms better classified are those belonging to normal class, while benign mammograms are the worst classified.

ROC curve is graphical display of sensitivity (TPR) on y-axis and (1 - specificity) (FPR) on x-axis with changing the decision threshold. This is generally depicted in a square box for convenience and its both axes are from 0 to 1. Figure 2 depicts the ROC curve for the proposed method. The area under the ROC curve is an important criterion for evaluating diagnostic performance. Usually it is referred as the Az index. Maximum Az = 1 and it means diagnostic test is perfect in differentiating diseased with non-diseased



(a) normal (b) benign (c) malign

Fig. 2: The ROC curve for the proposed method.

		Assigned Class		
		normal	malign	benign
3*Actual Class	normal	250	4	6
	malign	10	142	8
	benign	16	21	113

(a) SVM [16]

		Assigned Class		
		normal	malign	benign
3*Actual Class	normal	253	2	5
	malign	9	140	11
	benign	10	9	131

(b) RPRF

		Assigned Class		
		normal	malign	benign
3*Actual Class	normal	246	6	8
	malign	15	133	12
	benign	16	18	116

(c) RF [17]

		Assigned Class		
		normal	malign	benign
3*Actual Class	normal	240	9	11
	malign	16	130	14
	benign	15	22	113

(d) PNN [1]

Table 1: Confusion matrices showing classification error results for (a) SVM, (b) RPRF, (c) RF and (d) PNN Classifiers.

subjects. The proposed methodology yielded an area under the ROC curve of 0.943.

Table displays the numerical results from the experiments. Classification Accuracy represents the overall performance of a classifier. It indicates the percentage of correctly classified positive and negative cases from the total number of cases. Our model yielded a higher accuracy rate, with a mean of 91.93% compared to SVM (88.6%), PNN (84.74%) and RF (86.84%). Sensitivity, also known as recall rate, measures the proportion of positives correctly identified. The proposed methodology yielded a higher sensitivity rate, with a mean of 97.31% compared to SVM (96.15%), PNN (92.31%) and RF (94.62%). The specificity measure represents the proportion of negatives that are correctly identified. Our model has a specificity of 93.87%. F-measure is widely used to evaluate classification techniques. It is a common evaluation metrics that combines precision and recall into a single value. Our proposed yields F-measure of 0.9511 which is only 0.9328 for SVM, 0.9040 for PNN and 0.9162 for RF. The Brier score is a well-known evaluation measure for probabilistic classifiers. It measures the average squared deviation between predicted probabilities for a set of events and their outcomes. The lower the Brier score of a model the better the predictive performance. Our proposed has a small Brier score 0.1544, explaining the good results of classification for this dataset.

As a summary to these simulations, it can be observed that the classification efficiency of the proposed classifier is better than other classifiers, for the mammogram classification problem of the database considered for the study.

6. Conclusion

In this paper, we have developed a reliable probabilistic algorithm for the classification of masses in Mammograms. The proposed method has acceptable performance compared

	AC (%)	SE (%)	SP (%)	Az	F1	Prec	BS	MCC
SVM [16]	88.60	96.15	91.61	0.9646	0.9328	0.9058	0.2242	0.8747
PNN [1]	84.74	92.31	90.00	0.9131	0.9040	0.8856	0.3154	0.8209
RF [17]	86.84	94.62	90.00	0.9053	0.9162	0.8881	0.2441	0.8432
RPRF	91.93	97.31	93.87	0.9433	0.9511	0.9301	0.1544	0.9092

Table 2: Performance measures comparison.

to that obtained by the used comparison methods. In the future, we aim to refine our proposal for false-positive reduction. Furthermore, we would like to apply the proposed approach on other medical images where probabilistic predictions are of great importance.

References

[1] A. T. Azar and S. A. El-Said. Probabilistic neural network for breast cancer classification. *Neural Computing and Applications*, pages 1–15, 2012.

[2] L. Breiman and A. Cutler. Random forests - classification manual. *http://www.math.usu.edu/ãdele/forests/*, 2008.

[3] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth Inc, 1984.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] M. Dashevskiy and Z. Luo. Reliable probabilistic classification and its application to internet traffic. In *Advanced Intelligent Computing Theories and Applications*, volume 5226, pages 380–388, 2008.

[6] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M Parkin. Globocan 2008 v1.2. cancer incidence, mortality and prevalence worldwide in 2008. *IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer; 2010. [Accessed December 1, 2011]. at http://globocan.iarc.fr/*, 2008.

[7] M. J. Islam, M. Ahmadi, and M. A. Sid-Ahmed. Computer-aided detection and classification of masses in digitized mammograms using artificial neural network. *ICSI (2)'10*, pages 327–334, 2010.

[8] K. Kumar, P. Zhang, and B. Verma. Application of decision trees for mass classification in mammography. In *International conference on fuzzy systems and knowledge discovery, FSKD'06, China*, pages 366–376, 2006.

[9] P. Leod and B. Verma. Multi-cluster support vector machine classifier for the classification of suspicious areas in digital mammograms. *International Journal of Computational Intelligence and Applications*, 10(4):481– 494, 2011.

[10] J Liu, J Chen, X Liu, and J. Tang. An investigate of mass diagnosis in mammogram with random forest. In *Advanced Computational Intelligence (IWACI)*, pages 638 – 641, 2011.

[11] D. Meyer, F. Leisch, and K. Hornik. The support vector machine under test. *Neurocomputing*, 55(1-2):169–186, 2003.

[12] M. E. Osman, M. A. Wahed, A. S. Mohamed, and Y. M. Kadah. Computer aided diagnosis system for classification of microcalcifications in digital mammograms. In *26th National Radio Science Conference*, pages 1–6, 2009.

[13] H. Papadopoulos. Reliable probabilistic classification with neural networks. *Neurocomputing. Elsevier*, 2012.

[14] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, D. Betal, and J. Savage. The mammographic images analysis society digital mammogram database. *Experta Medica International Congress Series*, 1069:375–378, 1994.

[15] Y. Sun, C. F. Babbs, and E. J. Delp. Normal mammogram classification based on regional analysis. In *The 2002 45th Midwest Symposium on Circuits and Systems*, pages II–375 – II–378, 2002.

[16] G. vaira Suganthi and J. sutha. Classification of breast masses in mammograms using support vector machine. *IJCA Proceedings on International Conference on Recent Advances and Future Trends in Information Technology (iRAFIT 2012)*, iRAFIT(2):1–6, April 2012. Published by Foundation of Computer Science, New York, USA.

[17] L. Vibha, G. M. Harshavardhan, K. Pranaw, P. Deepa Shenoy, K. R. Venugopal, and Lalit M. Patnaik. Classification of mammograms using decision trees. In *Tenth International Database Engineering and Applications Symposium (IDEAS 2006), 11-14 December 2006, Delhi, India*, pages 263–266. IEEE Computer Society, 2006.

[18] V. Vovk, G. Alex, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. Springer, New York.

[19] C. Zhou, I. Nouretdinov, Z. Luo, M. Adamskiy, N. Coldham, and A. Gammerman. A comparison of venn machine with platt's method in probabilistic outputs. In *EANN/AIAI (2)'11*, pages 483–490, 2011.