Journal of Current Science and Technology, September-December 2022 Vol. 12 No. 3, pp. 391-407

ISSN 2630-0656 (Online)

Cite this article: Mathew, T. E. (2022, September). Appositeness of Hoeffding tree models for breast cancer classification. *Journal of Current Science and Technology, 12*(3), 391-407. DOI: 10.14456/jcst.2022.30

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Appositeness of Hoeffding tree models for breast cancer classification

Tina Elizabeth Mathew

Computer Science, Government College Kariavattom, Trivandrum, Kerala 695581, India

E-mail: tinamathew04@gmail.com

Received 28 April 2022; Revised 7 May 2022; Accepted 8 May 2022;
Published online 26 December 2022

Abstract

Supervised machine learning models have been shown to be effective in disease-related classification and prediction tasks by employing several classifiers. A prominent category among the set of supervised machine learners is decision trees. Decision Trees comprises of an assortment of tree classifiers. Each of these types of decision trees are extensively used as supervised learners for various classification problems. In this paper, to deal with the classification of breast cancer tumours into malignant or benign types, a subcategory of decision trees so called Hoeffding Trees are employed. Hoeffding Trees is a type of decision tree classifier that are usually effective when working with data streams. In this paper, we explore the performance and appropriateness of Hoeffding trees in building models to classify breast cancer tumours as either benign or malignant. Individual and ensemble models using Hoeffding trees are implemented for classification of breast cancer. In the work proposed here a class-balancer Hoeffding Tree model is realized and it was seen demonstrating the best performance among the different Hoeffding Tree models employed. The proposed model yielded an accuracy of 97.9%. Several other performance measures are also used to evaluate the performance of the implemented Hoeffding tree models. This work highlights the appositeness of Hoeffding tree models for breast cancer classification.

Keywords: *bagging; boosting; breast cancer (BC); class balancer (CB); decision tree (DT); ensemble; Hoeffding tree (HT).*

1. Introduction

Breast cancer otherwise known as the neoplasm of the breast is the cancer affecting the breast. The breast comprises of three main parts - the lobules, glands that produce milk, ducts, tubes that transport milk to the nipple and connective tissue, that surrounds and holds the elements together. Breast cancer occurs when there is uncontrolled cell growth, and is usually found, mostly, in the lobules and ducts of the breast. This uncontrolled growth causes a mass of cells termed as, tumours, to form. The tumours can be either benign or malignant. Malignant tumours are deadly and can spread the cancer throughout the body, whereas benign tumours are not problematic. Malignant tumours need to be identified and treated as early as possible to curb the spread of the cancer. Being prevalent in women and rare in men, breast cancer is presently considered to be the most commonly diagnosed cancer. Globocan (2020) reports that it has now surpassed lung cancer to be the one with most incidence. In 2020, there were an estimated 2.3 million new breast cancer cases, amounting to an 11.7% share of all cancers detected (Sung et al., 2021). Breast cancer remains the first or second leading cause of death in women before age 70 in 112 out of 183 countries and ranks third or fourth in another 23 countries, with a worldwide share of 6.9% of death cases. By 2040, the global cancer burden is expected to rise by 47% (Sung et al., 2021). Incidence rate in India is of no exception. Breast cancer comprises 34% of the total cancer cases in India (Sathishkumar et al., 2021) and still continues its upsurge.

The aforementioned statistics indicate the importance of, and the public health concern and risk posed by breast cancer. A major concern is the late detection and identification of the disease. This affects the treatment, survival prospects and life expectancy of the patient. Early detection is at most vital for better prognosis of the disease. Hence, it is imperative to build an infrastructure for dissemination of cancer identification and diagnostic measures. Early detection of cancer improves prognosis and aids in saving lives. In addition to conventional medical modalities and methods of cancer detection, computer-assisted techniques can be used as supplementary or alternatives for risk prediction (Vinod, & Manju, 2020), identification (Subash Chandra Bose, Sivanandam, & Praveen Sundar, 2021), prediction (Yadav, & Pal, 2019), diagnosis (Vidyapith, 2020) and classification (Seraphim, & Poovammal, 2021). Assistance utilizing these supplementary techniques will help medical practitioners besides, exempting the patients from the tiresome and painful process involved in the identification process of some medical diagnostic techniques.

Machine learning is a subdomain in the field of Computer Science that can support development of models for numerous purposes in various domains ranging from sentiment identification (Onan, 2019), text classification (Onan, Korukoğlu, & Bullut, 2016) sentiment analysis (Onan, 2020; Onan, 2021), sarcasm identification (Onan, 2021), mining of opinions (Onan, 2020), online fraud detection (Sailusha, Gnaneswar, Ramesh, & Rao, 2020), video surveillance systems (2020) to medical diagnostic and predictive purposes. In a recent work, Elen, and Avuchu (2021) highlighted the importance of machine learning techniques in analysing and reviewing problems in many areas. Boeri et al (2020) in their work were able to establish the importance of machine learning techniques in giving better prognosis as well, as a resource, an additional and inexpensive one, to aid identification, diagnosis and predict recurrence of the disease. In their comparative study (Rajamohana, Umamaheswari, Karunya, & Deepika, 2020) highlighted the important role machine learning techniques can play in breast cancer diagnosis and classification.

Classification algorithms – a type of machine learning algorithm with many variations – have been widely used for classification problems in disease prediction (Olayinka, & Chiemeke, 2019), among other domains (Sultana, & Jilani, 2021). In the breast cancer classification area, specifically, various techniques have been applied (Salama, Abdelhalim, & Zeid, 2012; Ummadi, Venkata Ramana Reddy, & Eswara Reddy, 2018), including logistic regression (Mathew, 2019), support vector machines (SVMs); (Mathew, 2019), *k*-nearest neighbours (*k*-NN); (Mathew, & Kumar, 2021) and decision trees (Hasan, Abu Bakar, Siraj, Sainin, & Hasan, 2015; Mathew, 2022). These classifiers, in combination with various feature selection techniques (Mathew, & Kumar, 2020; Mathew 2022) and other soft computing concepts, have been proven to improve the classification performance of machine learning models.

Specifically, decision tree models have been used for breast cancer classification and have produced promising results in classification (Kapoor, & Rani, 2015; Saraswat, & Singh, 2020). A review of the literature shows that Hoeffding trees (HTs) – a type of decision tree classifier – have been used for many problem areas, effectively including disease classification (Tekur, & Jain, 2018), assessment of student performance (Arundthathi, Glory Vijayaselvi, & Savithri, 2017), text classification (Melethadathil, Chellaiah, Nair, & Diwakar, 2015) and provision of fairness in decision-making situations (Zhang, & Zhao, 2020). These studies show the suitability of HTs for many classification tasks.

Alhayali, Ahmed, Mohialden, and Ali (2020) proposed an ensemble using HTs and a naïve Bayes classifier. The proposed model improved classification accuracy over the standard HT model on breast cancer data sets. Deepa, Senthil, and Singh (2019) showed that model performance can be enhanced when combining classifiers or various other techniques. Manju and Amrutha (2018) showcased the importance of feature selection with decision tree classifiers, including HTs, by demonstrating that feature selection improved performance. (Mathew, 2019) compared the performance of different decision tree classifiers in breast cancer classification and found that HTs performed best. (Tekur, & Jain, 2018) used HTs for colon cancer classification, with the HT classifier providing an accuracy of 90.3% against other classifiers such as naïve bayes, logistic regression and decision trees. Islam et al (2020) compared various machine learning techniques such as support vector machines, logistic

regression, k-nearest neighbours, artificial neural networks. ANN displayed the best performance but it was seen time consuming. The model performance based on building time is to be improved. Benbrahim, Hachimi and Amine (2020), compared 11 classifiers for breast cancer classification and among them neural networks achieved, the best accuracy of 96.49% exactness.

One major problem with medical datasets is the skewedness or imbalance of data. It poses a major challenge in supervised learning models (Onan, 2019). This has to be addressed while developing models. The positive class (class with disease) which constitutes the minority class in most of the datasets will be fewer in number and the negative class (class without disease) or majority class will have more instances in the dataset. This will lead to moderate to high predictive performance measures on the whole but the minority class will lack a reasonable representation. This will manifest in the model as misclassification of the classes, and in the case of minority class as misclassification of positive classes. The issue of misclassification is critical for medical diagnosis (Onan, 2019) and has to be taken into account. This is an area of interesting and important research in machine learning. Many studies do not consider this element and this is addressed in this study.

In this study, HTs are implemented and proposed as a model for classification. The proposed model is compared with the traditional simple as well as three ensemble of HT classifiers to highlight the superior performance of the proposed one over the traditional models. Ensembles have been recognized as a promising research field in machine learning (Onan, 2018). Ensembles have been used in combination with techniques such as feature selection to enhance performance (Onan, & Korukoğlu, 2017). In this study class balancing is introduced with the classifiers. The experimental results indicate the superior performance of the proposed model. `The predictive performance of supervised machine learning methods (such as support vector machines, logistic regression, and K-nearest neighbors) have also been examined. The analysis done indicates superior performance of the proposed model.

The main organization of the paper consists of 5 sections. The next section gives the objective of the study, Section 3 summarises the materials and methods, and Section 4 presents the results, followed by the conclusion.

2. Objective

The objective of this study is to evaluate the performance of HTs in breast cancer classification and identify the best HT model for breast cancer classification as well propose an effective model for Breast Cancer classification with reduced misclassification of classes, specifically the positive or minority class. Hoeffding Trees are decision tree classifiers usually applied for data streams, and are seen to be a less explored area in medical diagnostics. To fill the aforementioned gap, the predictive capability of Hoeffding Trees is considered, and this advantage is implemented in the breast cancer classification problem in the study. Initially, the performance of individual Hoeffding Tree models is evaluated. The possibility of improving accuracy using ensemble methods is explored. To ensure that skewness of data doesn't affect performance of the model class imbalance is considered and a model that implements this concept is developed.

A typical trend nowadays is to use deep learning models to solve classification problems. The most widely utilized deep learning models in various domains are CNN and RNN (Onan, 2022). Siddiqui et al (2021) in their study explores the breast cancer CAD method based on multimodal medical imaging and decision-based fusion., a deep learning approach was applied in multimodal medical imaging fusion, the resultant model obtained 97.5% accuracy with a 2.5% miss rate for breast cancer prediction. In their work, (Tiwari, Bharuka, Shah, & Lokare, 2020) compared machine learning techniques and deep learning techniques for prediction in breast cancer. Machine learning models svm, random forest gave accuracy of 96.5% and CNN gave an accuracy of 97,3%. Even though deep learning models show enhanced predictive performance they have a few drawbacks. The challenge faced by these models is that they require very large amount of data for better performance than traditional machine learning techniques. They need expensive GPUs and can be extremely expensive, besides producing complex data models. For smaller datasets machine learning techniques are considered effective. The hitch is to choose the most apt classifier for creating models. The algorithms of Hoeffding Trees have guaranteed high performance which is not common with the other decision tree learners (Moayedi, Jamali, Gibril, Foong, & Bahiraei, 2020), The study

proposed here investigates the appropriateness of Hoeffding trees as models for BC classification.

3. Methodology

3.1 Data set used

The Breast Cancer Wisconsin Data Set was used for this study. Created by Dr William H. Wohlberg of the University of Wisconsin Hospitals, Madison, the data set has 699 instances, 16 of which have missing values, and 11 predictor variables. For our analysis, we removed the instances with missing values and kept the remaining instances. The response variable is the class variable, which takes two values: 2 for benign tumours and 4 for malignant tumours. Nine predictor variables were selected. The first predictor variable (ID number) was omitted because it has no relevance in the study.

3.2 Techniques used

The various techniques used in the study are described in the following subsections.

3.2.1 Hoeffding trees

HTs are anytime decision trees that are capable of learning from massive data streams, provided that the distribution of the data is not altered over time. They make use of small sample sizes that are often adequate to choose the optimal splitting attribute. The usage of the Hoeffding bound helps in achieving this, as the Hoeffding bound can (within a prescribed precision) quantify the number of observations needed to estimate the suitability of the attribute. Hoeffding trees have two parts Hoeffding Tree training and Hoeffding Tree scoring. In Hoeffding tree training supervised learning is done to analyze a small sample data with known outcomes to choose the attribute for tree node splitting. In scoring it references the trained data and classifies each new instance into the concerned class label. Various studies have indicated that HTs demonstrate strong performance (Hulten, Spencer, & Domingos, 2001). The HT algorithm compares attributes better than other algorithms, in addition to having lower memory consumption and enhanced utilisation with sampling of data. However, it spends extensive time inspecting if ties occur (Deepa et al., 2019). HTs have been used for classification (Kumar, Kaur, & Sharma, 2015) and heart disease prediction (Benllarch, Benhaddi, & El Hadaj, 2021), where they have been shown to provide good performance. Hoeffding Trees being a category of Decision Trees are effective methods to handle nonlinear data. Tree-based classifiers have a non-linear and hierarchical approach that make them suitable for non-parametric as well as categorical data. They are seen to have a good deal of flexibility in data analysis and reveal the hierarchical structure of independent variables which is advantageous for classification. The algorithm of Hoeffding Trees is described below

Step 1: Select the provided data

Step 2: Every training data is filtered down incrementally to a suitable leaf.

Step 3: Every leaf node, say h, has enough data with it that is required to make decision about the next step. This data at leaf node estimates the information gain (Attribute Selection criteria) when any attribute is split.

Step 4: the best attribute at a node is to be found and a test based on provided data, to decide whether a particular attribute has produced better result than other attributes using Hoeffding bound, is to be performed.

Step 5: After applying a number of tests, the attribute, which provide better result than any other node, results in splitting the node for growth of tree

3.2.2 Ensembles

Instead of using single or individual classifiers alone, ensembles can be used to improve performance. Ensemble learning is a machine learning technique that improves the produced output using a combination of methods. Ensemble learning also termed as multiple classifier systems is defined as the process of training multiple learning algorithms and combining their predictions by regarding them as a group of decision makers abetting as predictive performance enhancers. (Onan, 2020). They have been used in a wide range of domains including topic extraction (Onan, 2019), sentiment analysis (Onan, & Korukoğlu, 2017) and medical and biomedical applications (Onan, 2018). Ensembles can be categorised as homogenous or heterogeneous. Bagging, boosting and dagging ensembles are considered to be homogenous ensembles (Onan, Korukoğlu, & Bulut, 2017), as they involve the same type of classifiers. In heterogeneous ensembles, different types of classifiers can be combined (Shastri et al., 2021). Stacking ensembles belong to the category of heterogenous ensembles. Homogenous and heterogenous ensembles are seen to give better

performance than single classifiers (Baruah, Goswami, Bora, & Baruah, 2022; Phua, & Batcha, 2020). Ensemble learning schemes pursue identification of more robust classification with a better predictive performance (Onan, 2019). In his work (Onan, 2015) indicates that ensemble learning can improve the predictive performance of base learners in medical domain. In the preliminary approach used in this study, only homogenous ensembles are implemented.

3.2.2.1 Bagging

Bagging, or bootstrap aggregating, is an ensemble method that is considered more robust than using individual learners. The concept involved in this method helps to reduce variance and prevent overfitting issues. Consider a training set $S\{(y_n; x_n), n = 1,\ldots., N\}$, where y is a class or numerical target response and x is the input. Samples S_B are created from this set with replacement. The classifier is applied on each of these samples, and the results obtained are aggregated using majority vote. (Ponnaganti, & Anita, 2022) used bagging ensembles to improve accuracy of their proposed work. It was seen to provide effective training to machine learning classifiers with reduced computational time along with improved performance. Bagging was also seen to overcome class imbalance problem as well as increase accuracy measure as analyzed by (Saputra, & Prasteyo, 2020).

3.2.2.2 Boosting

Boosting is another ensemble technique. Boosting is similar to bagging except that it fits the samples to the classifier in a sequential manner. The advantage of boosting over bagging is that the misclassification that occurs in one run can be overcome by assigning more weights to the misclassified instances in the next run. The focus of boosting models is to reduce bias. (Osman, & Aljadhali , 2020) were able to produce better performance in machine learning classifiers used for breast cancer prediction by implementing ensemble boosting. These ensemble strategies have proven to be useful in modelling complex, heterogeneous datasets of any size such as in radiomics research. (Vamvakas, Tsivaka, Logothetis, Vassiou, & Tsougos, 2022).

3.2.2.3 Dagging

Dagging forms a number of disjoint, stratified folds out of the data and feeds each of these sets of data to a copy of the specified base classifier. The dagging metaclassifier integrates all the results and uses majority voting to provide the final output. Consider a training dataset containing n samples. k subsets are constructed by randomly taking samples from the dataset without replacement such that each of them contains say, n' samples, where $kn' \leq n$. A chosen classifier is trained on these k subsets, producing k classification models M_1, M_2, …, M_k. For any test input, $M_i(1 \leq i \leq k)$ provides an output and the final predicted result of dagging model is the class with most votes.

3.2.3 Class balancing

Class balancing can be done using a few techniques. Class balancing techniques can be categorised into two major groups: data-based techniques and algorithm-based techniques. Data pre-processing can be employed to tackle the class imbalance issue (Onan, 2019). Each of these categories of class balancing Techniques are further divided into subcategories. Class-balancer techniques are a subcategory of data-based techniques. Data-level approaches have a greater potential over imbalanced learning as they aid to improve the distribution of datasets, rather than relying on enhancements made on supervised learning techniques (Onan, 2019; Barua, Islam, Yao, & Morase, 2012). In class balancing, each class of instances is assigned instance weights so that they have the same weight and the total sum of instance weights of the data set is unaffected. In this model this class balancer technique is used.

3.2.3.1 Proposed CB-HT model

The methodology of the proposed class-balancer–HT (CB-HT) model is illustrated in Figure 1. Initially, the data set is preprocessed. The missing values and irrelevant attributes are removed as part of preprocessing. The data set is then split into two sets for training and testing, respectively, using an 80–20 partition. Class balancing is performed on the training data set. The HT classifier is then applied on the class-balanced training partition, and the model is trained. To avoid overfitting, 10-fold cross-validation is used. Later, the model is also run on the testing set. To compare performance of the proposed model, the data set is used to train individual and ensemble HT classifier models.

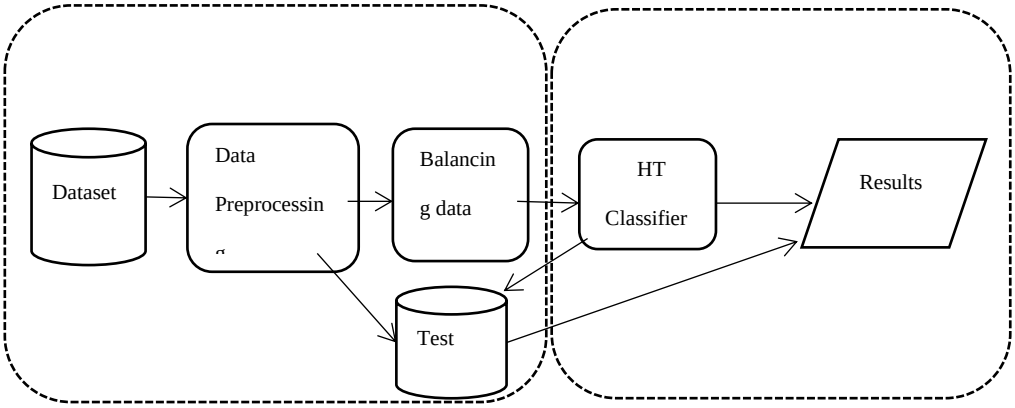

Figure 1 Working of proposed model

The proposed CB-HT model is also compared with these HT models. Initially, a simple HT model is created, and the data set is applied on this model. The performance of ensemble HT models in breast cancer classification was also explored. HT with bagging, HT with boosting and HT with dagging ensemble models were created. In addition to illustrating better classification accuracy, the model aims to reduce cases with type I and type II errors.

4. Results and discussion

The performances of the simple ensemble and proposed CB-HT model on the Wisconsin breast cancer dataset are shown in Table 1 and Table 2, respectively in terms of Accuracy, Kappa Statistic, F measure, MCC, Precision, Recall, FPR, ROC, and PR AUC. The dataset contains samples from the two classes malignant and benign. The results show that the proposed model performs well with the dataset in classification. For medical applications accuracy is not a completely dependable measure, Lu et al (2015) hence other measures are used for evaluation. The HT classifier obtained an accuracy of 97.36%, a kappa statistic value of 0.9425, a Matthews correlation coefficient (MCC) of 0.943 and a receiver operating characteristic curve (ROC) of 0.994. The bagging ensemble model had better accuracy (97.51%) as well as better kappa and MCC values (0.9458 and 0.946, respectively). The misclassification of the bagging model slightly improved on that of the simple HT model. In the boosting model, accuracy was reduced to 96.9%, and kappa and MCC values were relatively lower at 0.9327 and 0.933, respectively. Misclassification of the malignant class was seen to be higher for the boosting model than the bagging model. Among the ensemble models, the dagging model exhibited the worst performance with 96.77% accuracy, a kappa value of 0.9293, and an MCC value of 0.929. The dagging model's misclassification of the malignant class and overall time taken were both higher compared with the other models. Besides the time taken to build by the ensemble models are seen to be more than the proposed and simple HT models. A plausible reason for the poor performance of the boosting model may be that the model over-emphasizes instances that are noise as a result of overfitting of the training data set. This could also be the reason the bagging model was seen to perform better. Boosting models are also sensitive to outliers whereas bagging is robust to outliers. Besides, bagging models handle irrelevant features well even without feature scaling. Ensembling performs well when moderately performing classifiers are combined.

The proposed model (i.e., the CB-HT model) demonstrated a better accuracy of 97.9% and a better *F*-measure of 0.979. The kappa and MCC values obtained were 0.9582 and 0.958, respectively. Misclassification of both benign and malignant classes was reduced. This indicates that the CB-HT model had comparatively better performance. The proposed model had a false positive rate (FPR) of 0.021 and a true positive rate (TPR) of 0.979. These results indicate the appropriateness of the proposed HT model for breast cancer classification.

Table 1 Performance of simple and ensemble Hoeffding trees

Classifiers	Hoeffding Tree	Bagging-Hoeffding Tree	Boosting-Hoeffding Tree	Dagging-Hoeffding Tree
Accuracy Percentage	97.36	97.51	96.9	96.77
Kappa Statistic	0.9425	0.9458	0.9327	0.9293
Precision	0.974	0.976	0.97	0.968
Recall	0.974	0.975	0.969	0.968
MCC	0.943	0.946	0.933	0.929
F-Measure	0.974	0.975	0.969	0.968
ROC	0.994	0.993	0.978	0.994
P-R AUC	0.993	0.998	0.971	0.993
FPR	0.024	0.021	0.032	0.037
Time (sec)	0.01	0.03	0.03	0.06
Confusion Matrix	431 13 5 234	431 13 4 235	431 13 8 231	432 12 10 229

The confusion matrix is a table that assists evaluating the performance of the classifier Each row in the confusion matrix represents the actual class, while each column represents the predicted class. It also shows the misclassification of instances in the off-diagonal cells. Amongst the various models here, the bagging model has the best values for lower misclassification of the positive class among the models as in Table 1, and the dagging model has the worst number of misclassified instances. The false positives produced by the different models are high in all cases. The proposed CB-HT model had lesser misclassification of true positives (4 instances) and lesser misclassification of true negatives (10 instances) than shown by other HT models.

Table 2 Performance of proposed model

Model Used	Proposed CB-HT Classifier
TPR	0.979
FPR	0.021
Accuracy	97.9
Kappa Statistic	0.9582
Precision	0.979
Recall	0.979
F- Measure	0.979
MCC	0.958
ROC Area	0.993
PRC	0.993
Confusion Matrix	332 10 4 337
Time (in secs)	0.01

Figure 2 depicts the rates of incorrect classification for the models. The proposed CB-HT model had the least incorrect classification at 2.0916%, and the dagging model demonstrated the worst classification rate at 3.2211%. The CB-HT model obtained a comparatively better reduction of false positives and false negatives. Recall of the CB-HT model was obtained as 0.979. Recall, generally referred to as sensitivity (Equation (4)), is ratio of the positive predictions that are effectively predicted as positive. This measure is significant, specifically within the clinical field as it indicates the level of the observations that are accurately analyzed during experimentation. It is imperative to appropriately recognize a threatening neoplasm than inaccurately distinguishing a less problematic one. Likewise, precision of the CB-HT model was obtained as 0.979. Precision illustrates how well the classifier handles the positive observations however it doesn't say much regarding the negative observations. F1 score of the CB-HT model is likewise, 0.979. F1-Score is defined as the weighted harmonic mean of Precision and Recall. The measure considers both false positive and false negative instances.

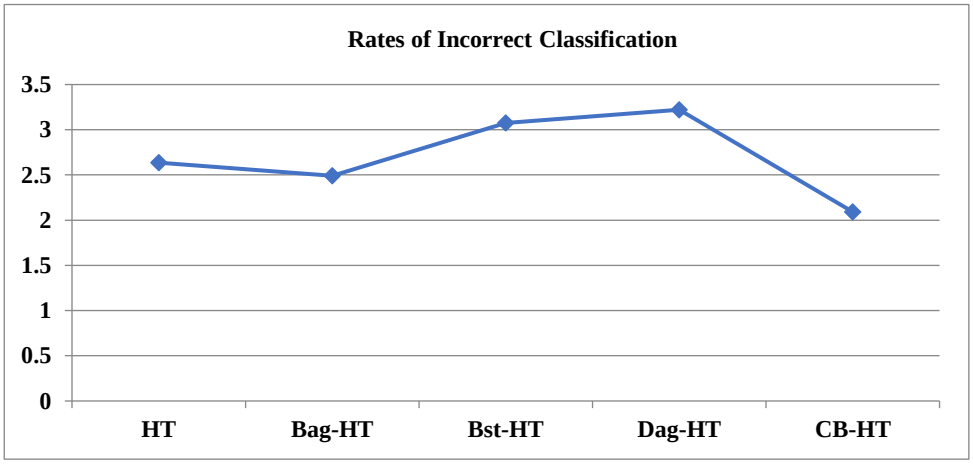

Figure 2 Incorrect classification rates

Figure 3 illustrates a representation of the various performance evaluation measures used. The CB-HT model has better performance measures in most cases, though the bagging model performed best in terms of the area under the precision-recall curve metric. In particular, the MCC value of the CB-HT model is much higher compared with the other models. MCC is a balanced measure and can be used even when the classes have different sizes. It is often used to assess performance of prediction methods in a two-class classification problem This metric depicts the quality of the binary classifier and is considered more reliable than the accuracy measure. (Gu, Zhu, & Cai, 2009). Data sampling is considered as a consistent solution for data imbalance as it modifies the structure of actual data-set by altering the balance ratio. This has been proven as consistent and a significant factor for classification enhancement in this work.

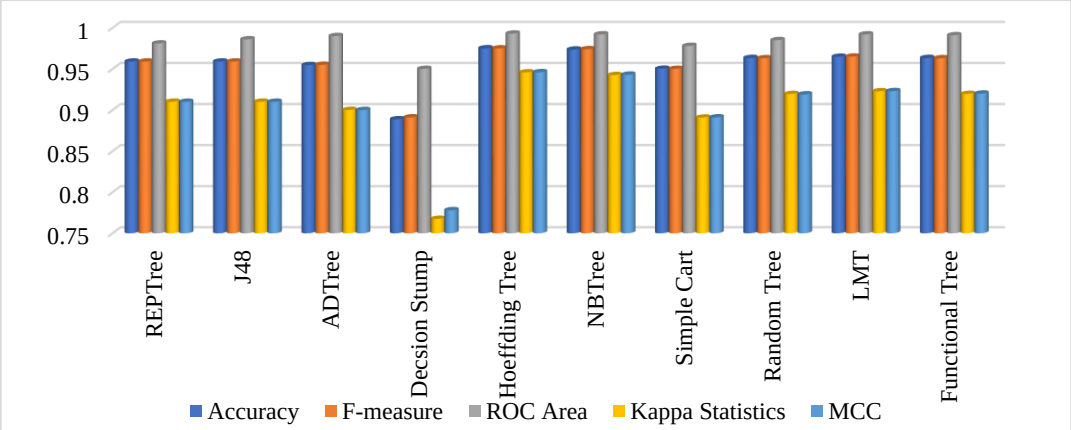

Figure 3 Comparison of performance measures

Figure 4 represents the FPR of each model and the time taken to build the various HT models. The bagging model had the lowest FPR, and the dagging model took the most time to build. Time taken by the bagging and boosting models were same as in Table 1but higher than the individual model. This is because each classifier works independently in bagging and sequentially in boosting. The proposed model required lesser time than the ensemble models.

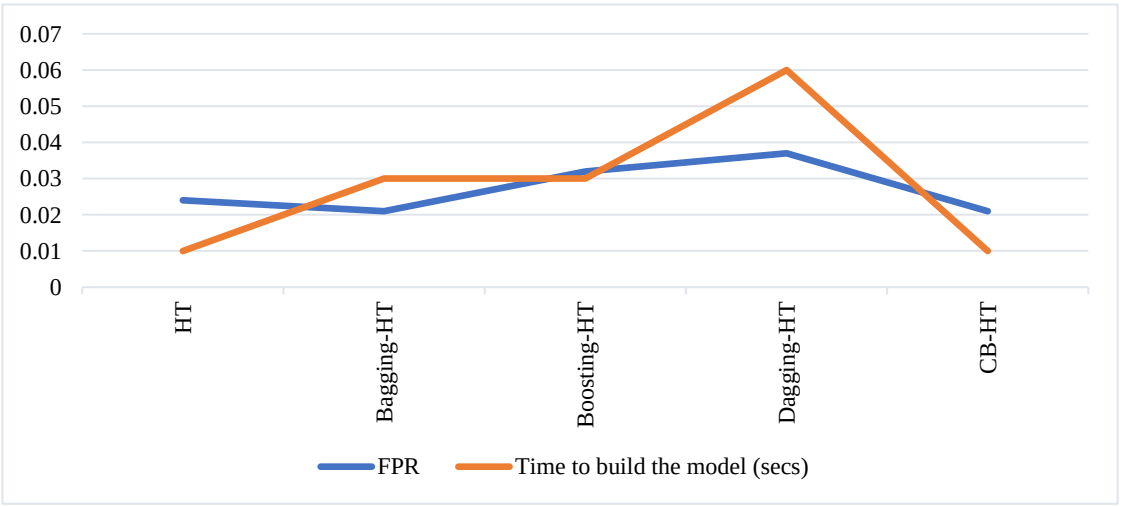

Figure 4 False positive rates and time taken to build the models

Accuracy is defined as classifier exactness, is a proportion of how well the classifier can accurately predict or classify instances into their right classification. As in Equation1. It's the number of correct forecasts separated by the whole number of instances within the data set. For balanced datasets, accuracy is an appropriate measure. Accuracy is measured as in Equation (1). Since the data set is imbalanced for the simple and ensemble HT models, the accuracy measure alone can be misleading. Hence, measures such as balanced accuracy, as in Equation (2) and geometric mean, as in Equation (3), is used for the simple and ensemble HT models.

Accuracy = (TP+TN)/(TP+TN+FP+FN) Equation (1)

Balanced Accuracy = (Sensitivity+Specificity)/2 Equation (2)

Geometric Mean = √((Sensitivity*Specificity) Equation (3)

where, Sensitivity = TP/(TP+FN) Equation (4)

and, Specificity = TN/(TN+FP) Equation (5)

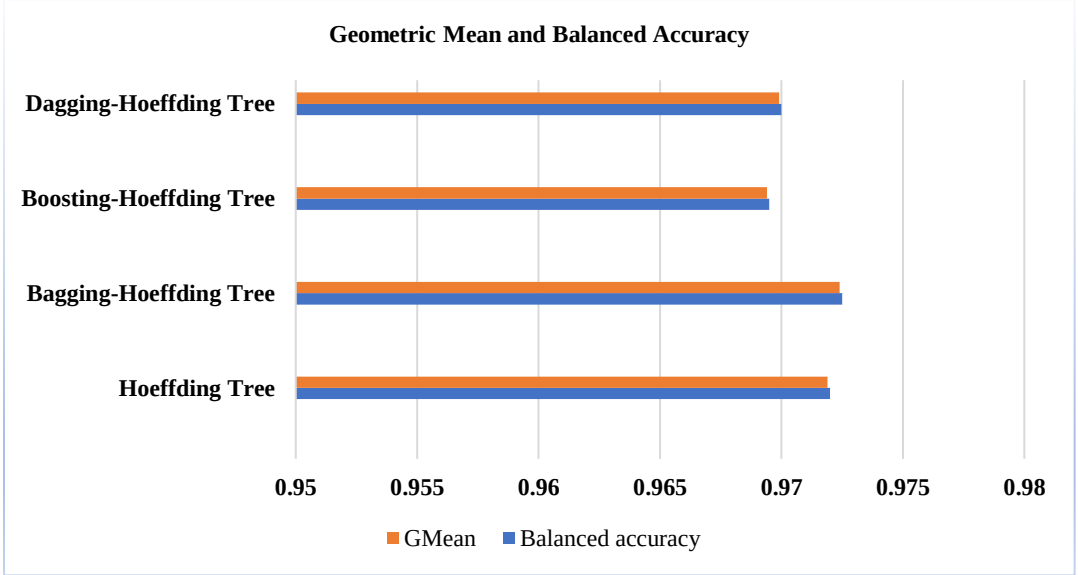

Figure 5 Geometric mean and balanced accuracy

In all cases the proposed CB-HT model outperformed the other models in accuracy. Figure 5 depicts the results for these two performance measures. With these measures, it is obvious that the bagging model performs much better compared with the other simple and ensemble models. Geometric mean is the geometric mean of sensitivity and specificity, while balanced accuracy is the average of the sum of specificity and sensitivity. The geometric mean and balanced accuracy values obtained by the models are almost the same. When dealing with imbalanced datasets, GM and BM are the best performance metrics if their focus is better success. However, if classification errors are to be considered, then MCC will be a better choice (Luque, Carrasco, Martin, & de las Heras, 2019). Based on MCC the bagging model is better among the individual and ensemble models.

The proposed model is compared with other Decision tree models and this is illustrated in figure 6. The figure displays the accuracy of each type of decision tree and that of the proposed model. The proposed model displayed better performance than all other classifiers

The proposed CB-HT model is also compared with few other state of art machine learning classifiers such as Support Vector Machines (SVM), Logistic Regression (LR) and k-Nearest Neighbour (k-NN) classifiers and figure 7 displays the superior performance of the proposed model. Classification accuracy is taken as the metric for evaluation of the proposed model with these techniques. Comparison results demonstrate that the performance of the proposed approach is significant compared to other competing techniques. The LR classifier gave an accuracy of 92.53%, the SVM gave accuracy of 96.19, and k-NN gave accuracy of 94. 8% against 97.9% of the proposed CB-HT model. The misclassification of the proposed model was far less than that of the three other classifiers. The methods and techniques implemented in this proposed work is more advantageous compared to the existing methods as it improves and predicts the tumor at its earliest stages. Overall, the proposed model proved to be efficient in detecting benign and malignant class labels which was evident through the statistical analysis of the models. However, the scope of the proposed research is limited to a specific dataset. In future the scope of this work is to be expanded by conducting experiments with more extensive data sets.

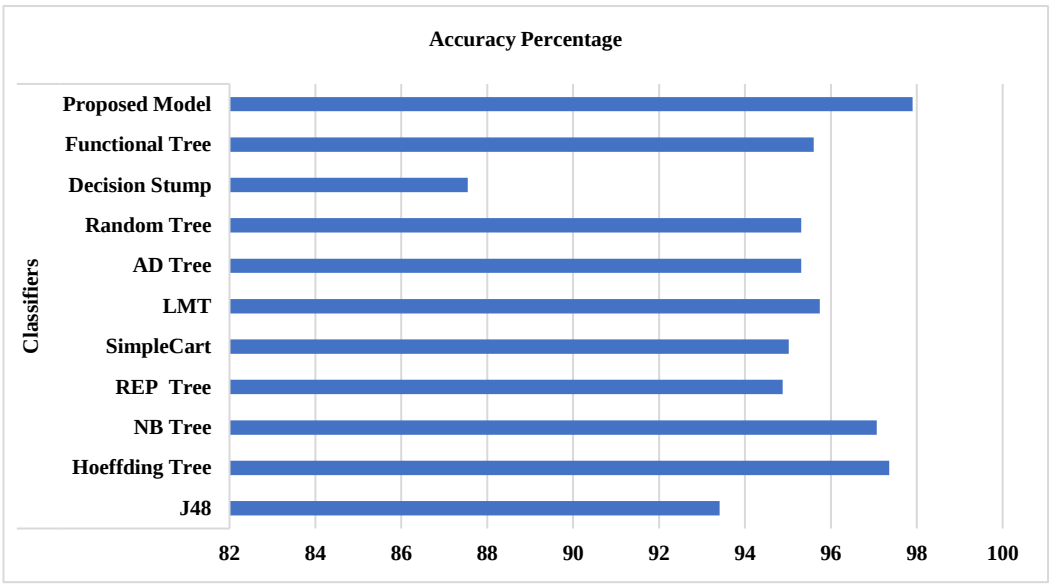

Figure 6 Proposed model vs decision trees

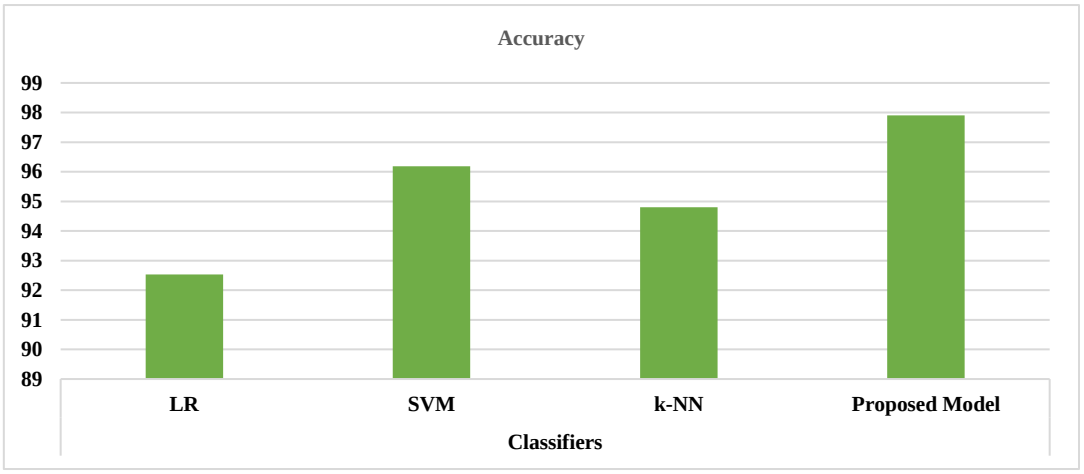

Figure 7 Proposed model vs other classifiers

The proposed model is compared with some state of art technologies. Chaurasia, Pal and Tiwari (2018) in their proposed work suggested Naïve Bayes as the best predictor with 97.36% accuracy (this prediction accuracy is better than any reported in the literature), RBF Network came out to be the second with 96.77% accuracy, J48 came out third with 93.41% accuracy. (Zheng et al., 2020) proposed a Deep Learning assisted Efficient Adaboost Algorithm (DLA-EABA) for breast cancer diagnosis and the model produced an accuracy of 97.2%. Jabbar (2021) proposed an ensemble model built with Bayesian network and Radial Basis Function and obtained an accuracy of 97% on the WBC Dataset. Imran et al (2022) in their proposed work compared Random Forest, Naïve Bayes and AdaBoost for breast cancer classification using many sets varying the values of k in k fold cross validation, 10-fold obtained a result of 96%, 94% and 92% accuracy respectively. In all cases the proposed model is seen to produce a better performance. The comparison is seen in Table 3.

Table 3 Comparison with existing literature

Author	Classifier used	Accuracy
Chaurasia et al (2018)	Naïve bayes	97.36
	RBF Network	96.77
	J48	93.41
Zheng et al (2020)	DLA-EABA	97.2
Jabbar (2021)	Bayesian network- RBF	97
Imran et al (2022)	Random Forest	96
	Naïve Bayes	94
	AdaBoost	92
Proposed Model	Hoeffding Trees	97.9

5. Conclusion

In this study, the intention was to identify the best Hoeffding Tree model for breast cancer classification as well as to develop a model that reduces misclassification, specifically, of the minority class also with, better classification accuracy. The key contribution is a class balancer Hoeffding Tree (CB-HT) model. The proposed model was able to attain both of these objectives. The issues related with class imbalance is addressed. In the paper, various categories of simple and ensemble techniques using HT classifiers were examined. It was found that the proposed CB-HT model demonstrated the best classification compared with the various other techniques used and had comparatively lower misclassification rates. An accuracy score of 97.9% was obtained by the proposed model. The proposed model was compared with individual and homogenous ensemble classifiers. The proposed model was seen superior in performance. A major issue seen with models is even with high accuracy value misclassification will still exist. Misclassification of the positive class which is usually the minority class is a much more serious matter than misclassification of the negative class. This issue was significantly reduced in this model. Other Hoeffding Tree models were also compared. Amongst the ensemble models, the bagging ensembles were seen to work better than the other two ensemble techniques explored in the experiments (i.e., boosting and dagging). The model is also compared with other categories of decision tree classifiers. Amongst them the proposed model produced superior performance with better accuracy and lesser misclassification. Comparison with other categories of classifiers - SVM, LR, k-NN also highlighted the superior classification of the proposed model. The results obtained in this study highlights the appropriateness of HT models for breast cancer classification. The work of this kind, intends to contribute to the efforts of developing an alternative in breast cancer classification so as to provide support in the examination process of a patient by implementing a painless, non-invasive and harmless technique, in addition to diagnosing breast cancer effectively. The work suggests the same can be achieved by this model.

Future work could investigate the performance of other categories of ensembles – specifically heterogenous ensembles, such as stacking ensembles – with the HT classifier. Further techniques, such as optimization techniques, feature selection, could also be incorporated into the models to further improve the outcomes. An alternative approach for class balancing using algorithm-based techniques can also be explored. The proposed work is implemented on a small dataset of size 699 x 9. This can be extended for larger datasets. Besides, the performance of the model on other datasets from different domains can be investigated and compared. In the proposed model the rate of false negatives was much reduced but that of false positives needs further reduction. Techniques to improve this can be implemented.

6. Acknowledgements

I would like to thank Dr William H. Wolberg of the University of Wisconsin Hospitals, Madison, for the data set.

7. References

Alhayali, R. A. I., Ahmed, M. A., Mohialden, Y. M., & Ali, A. H. (2020). Efficient method for breast cancer classification based on ensemble hoffeding tree and naïve Bayes. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(2), 1074-1080. DOI: 10.11591/ijeecs. v18.i2

Arundthathi, A., Glory Vijayaselvi, K., & Savithri, V. (2017). Assessment of Decision Tree Algorithm on Student's Recital. *International Research Journal of Engineering and Technology*, *4*(3), 2342-2348.

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2012). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, *26*(2), 405-425. https://doi.org/10.1109/TKDE.2012.232

Baruah, A. J., Goswami, J., Bora, D. J., & Baruah, S. (2022). A Comparative Research of Different Classification Algorithms. In *Intelligent Sustainable Systems* (pp. 631-646). Singapore: Springer. DOI: https://doi.org/10.1007/978-981-16-2422-3_50

Benbrahim, H., Hachimi, H., & Amine, A. (2019). Comparative study of machine learning algorithms using the breast cancer dataset. In *International Conference on Advanced Intelligent Systems for Sustainable Development* (pp. 83-91). Cham: Springer. https://doi.org/10.1007/978-3-030-36664-3_10

Benllarch, M., Benhaddi, M., & El Hadaj, S. (2021). Enhanced Hoeffding Anytime Tree: A Real-time Algorithm for Early Prediction of Heart Disease. *International Journal on Artificial Intelligence Tools*, *30*(03), 2150010. https://doi.org/10.1142/S021821302150010X

Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer medicine*, *9*(9), 3234-3243. DOI: 10.1002/cam4.2811

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119-126. DOI: doi:10.1177/1748301818756225

Deepa, B. G., Senthil, S., & Singh, P. (2019). Data Mining on Classifiers Prophecy of Breast Cancer Tissues. *International Journal of Advanced Networking and Applications*, *10*(5), 8-12.

Elen, A., & Avuçlu, E. (2021). Standardized Variable Distances: A distance-based machine learning method. *Applied Soft Computing*, *98*, 106855. https://doi.org/10.1016/j.asoc.2020.106855

Elhoseny, M. (2020). Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems. *Circuits, Systems, and Signal Processing*, *39*(2), 611-630.

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *International symposium on intelligence computation and applications* (pp. 461-471). Springer, Berlin, Heidelberg.

Hasan, M. R., Abu Bakar, N. A., Siraj, F., Sainin, M. S., & Hasan, S. (2015). *Single decision tree classifiers' accuracy on medical data*. Retrieved form https://repo.uum.edu.my/id/eprint/15527.

Hulten, G., Spencer, L., & Domingos, P. (2001, August). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97-106), https://doi.org/10.1145/502512.502529

Imran, B., Hambali, H., Subki, A., Zaeniah, Z., Yani, A., & Alfian, M. R. (2022). Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer. *Jurnal Pilar Nusa Mandiri*, *18*(1), 37-46. https://doi.org/10.33480/pilar.v18i1.2912

Islam, M., Haque, M., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 1-14, https://doi.org/10.1007/s42979-020-00305-w

Jabbar, M. A. (2021). Breast cancer data classification using ensemble machine learning. *Engineering and Applied Science Research*, *48*(1), 65-72.

Kapoor, P., & Rani, R. (2015). A Survey of Classification Methods Utilizing Decision Trees. *International Journal of Engineering Trends and Technology*, *22*(4), 188-194. DOI: 10.14445/22315381/IJETT-V22P240

Kumar, A., Kaur, P., & Sharma, P. (2015). A survey on Hoeffding tree stream data classification algorithms. *CPUH-Research Journal*, 1(2), 28-32.

Lu, J., Hales, A., Rew, D., Keech, M., Fröhlingsdorf, C., Mills-Mullett, A., & Wette, C. (2015, September). Data mining techniques in health informatics: a case study from breast cancer research. In *International Conference on Information Technology in Bio-and Medical Informatics* (pp. 56-70). Springer, Cham. https://doi.org/10.1007/978-3-319-22741-2_6

Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216-231, https://doi.org/10.1016/j.patcog.2019.02.023

Manju, B. R., & Amrutha, V. S. (2018). Comparative study of datamining algorithms for Diagnostic Mammograms using Principal component analysis and J48. *ARPN Journal of Engineering and Applied Sciences*, *15*(3), 354-362.

Mathew, T. E. (2019a). A comparative study of the performance of different Support Vector machine Kernels in Breast Cancer Diagnosis. *International Journal of Information and Computing Science*, *6*(6), 432-441. DOI: 16.10089/IJICS

Mathew, T. E. (2019b). A logistic regression with recursive feature elimination model for breast cancer diagnosis. *International Journal on Emerging Technologies*, *10*(3), 55-63.

Mathew, T. E. (2019c). Simple and ensemble decision tree classifier based detection of breast cancer. *International Journal of Scientific & Technology Research*, *8*(11), 1628-1637.

Mathew, T. E., & Kumar, K. A. (2020). A Logistic Regression based hybrid model for Breast Cancer Classification. *Indian Journal of Computer Science and Engineering (IJCSE)*, *11*(6), 899-903. DOI: 10.21817/indjcse/2020/v11i6/201106201

Mathew, T. E., Kumar, K. S., (2021). A Modified-Weighted- K -Nearest Neighbour and Cuckoo Search Hybrid Model for Breast Cancer Classification. *Indian Journal of Computer Science and Engineering (IJCSE)*, 12(1), 166-177. DOI: 10.21817/indjcse/2021/v12i1/211201211

Mathew, T. E. (2022a). An Improvised Random Forest Model for Breast Cancer Classification. *NeuroQuantology*, *20*(5), 713-722.

Mathew, T. E. (2022b). An Optimized Extremely Randomized Tree Model For Breast Cancer Classification. *Journal of Theoretical and Applied Information Technology*, *100*(16), 5234-5246.

Melethadathil, N., Chellaiah, P., Nair, B., & Diwakar, S. (2015, August). Classification and clustering for neuroinformatics: Assessing the efficacy on reverse-mapped NeuroNLP data using standard ML techniques. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1065-1070). IEEE, 978-1-4799-8792-4/15/$31.00

Moayedi, H., Jamali, A., Gibril, M. B. A., Kok Foong, L., & Bahiraei, M. (2020). Evaluation of tree-base data mining algorithms in land used/land cover mapping in a semi-arid environment through Landsat 8 OLI image; Shiraz, Iran. *Geomatics, Natural Hazards and Risk*, *11*(1), 724-741, DOI: 10.1080/19475705.2020.1745902

Olayinka, T. C., & Chiemeke, S. C. (2019). Predicting paediatric malaria occurrence using classification algorithm in data mining. *Journal of Advances in Mathematics and Computer Science*, *31*(4), 1-10. DOI: 10.9734/JAMCS/2019/v31i430118

Onan, A. (2015). On the performance of ensemble learning for automated diagnosis of breast cancer. In *Artificial intelligence perspectives and applications* (pp. 119-129). Springer, Cham, https://doi.org/10.1007/978-3-319-18476-0_13

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232-247., https://doi.org/10.1016/j.eswa.2016.03.045

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38, https://doi.org/10.1177%2F0165551515613226

Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, *53*(4), 814-833., https://doi.org/10.1016/j.ipm.2017.02.008

Onan, A. (2018a). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28-47, https://doi.org/10.1177%2F0165551516677911

Onan, A. (2018b). Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Computational and Mathematical Methods in Medicine*, *2018*, 1-22. https://doi.org/10.1155/2018/2497471

Onan, A. (2019a). Topic-enriched word embeddings for sarcasm identification. In *Computer Science On-line Conference* (pp. 293-304). Springer, Cham, https://doi.org/10.1007/978-3-030-19807-7_29

Onan, A. (2019b). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, *7*, 145614-145633., DOI: 10.1109/ACCESS.2019.2945911

Onan, A. (2019c). Consensus clustering-based under sampling approach to imbalanced learning. Scientific Programming, *2019*, 1-14. https://doi.org/10.1155/2019/5901087

Onan, A. (2020). Mining opinions from instructor evaluation reviews: a deep learning approach. *Computer Applications in Engineering Education*, *28*(1), 117-138, https://doi.org/10.1002/cae.22179

Onan, A. (2021a). Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Computer Applications in Engineering Education*, *29*(3), 572-589., https://doi.org/10.1002/cae.22253

Onan, A. (2021b). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), e5909., https://doi.org/10.1002/cpe.5909

Onan, A., & Toçoğlu, M. A. (2021). A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access*, *9*, 7701-7722., https://doi.org/10.1109/ACCESS.2021.3049734

Onan, A. (2022). Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 2098-2117. https://doi.org/10.1016/j.jksuci.2022.02.025

Osman, A. H., & Aljahdali, H. M. A. (2020). An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model. *IEEE Access*, *8*, 39165-39174.

Phua, E. J., & Batcha, N. K. (2020). Comparative analysis of ensemble algorithms'

prediction accuracies in education data mining. *Journal of Critical Review*, *7*(3), 37-40. http://dx.doi.org/10.31838/jcr.07.03.06

Ponnaganti, N. D., & Anitha, R. (2022). A Novel Ensemble Bagging Classification Method for Breast Cancer Classification Using Machine Learning Techniques. *Traitement du Signal*, *39*(1), 229-237.

Rajamohana, S. P., Umamaheswari, K., Karunya, K., & Deepika, R. (2020). Analysis of classification algorithms for breast cancer prediction. In *Data Management, Analytics and Innovation* (pp. 517-528). Springer, Singapore, https://doi.org/10.1007/978-981-32-9949-8_36

Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020, May). Credit card fraud detection using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1264-1270). IEEE, https://doi.org/10.1109/ICICCS48265.2020.9121114

Salama, G. I., Abdelhalim, M. B., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technolog*, *1*(1), 36-43.

Saputra, R. H., & Prasetyo, B. (2020). Improve the accuracy of c4. 5 algorithm using particle swarm optimization (pso) feature selection and bagging technique in breast cancer diagnosis. *Journal of Soft Computing Exploration*, *1*(1), 47-55, https://doi.org/10.52465/joscex.v1i1.9

Saraswat, D., & Singh, P. (2020). Comparison of Different Decision Tree Algorithms for Predicting the Heart Disease. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences* (pp. 245-255). Springer, Singapore, https://doi.org/10.1007/978-981-15-6318-8_21

Sathishkumar, K., Vinodh, N., Badwe, R. A., Deo, S. V. S., Manoharan, N., Malik, R., ... & Mathur, P. (2021). Trends in breast and cervical cancer in India under National Cancer Registry Programme: an age-period-cohort analysis. *Cancer Epidemiology*, *74*, 101982, https://doi.org/10.1016/j.canep.2021.101982

Seraphim, B. I., & Poovammal, E. (2021). Based Data Classification Techniques in Healthcare Using Massive Online Analysis Framework. *Machine Learning and Analytics in Healthcare Systems: Principles and Applications* (213). US: CRC Press.

Shastri, S., Kour, P., Kumar, S., Singh, K., Sharma, A., & Mansotra, V. (2021). A nested stacking ensemble model for predicting districts with high and low maternal mortality ratio (MMR) in India. *International Journal of Information Technology*, *13*(2), 433-446. https://doi.org/10.1007/s41870-020-00560-3

Siddiqui, S. Y., Naseer, I., Khan, M. A., Mushtaq, M. F., Naqvi, R. A., Hussain, D., & Haider, A. (2021). Intelligent breast cancer prediction empowered with fusion and deep learning, *Computers, Materials and Continua*, *67*(1), 1033-1049. http://dx.doi.org/10.32604/cmc.2021.013952

Subash Chandra Bose, S., Sivanandam, N., & Praveen Sundar, P. V. (2021). Design of ensemble classifier using Statistical Gradient and Dynamic Weight LogitBoost for malicious tumor detection. *Journal of Ambient Intelligence and Humanized Computing*, *12*(6), 6713-6723. https://doi.org/10.1007/s12652-020-02295-2

Sultana, J., & Jilani, A. K. (2021). Classifying Cyberattacks Amid Covid-19 Using Support Vector Machine. In *Security Incidents & Response Against Cyber Attacks* (pp. 161-175). Springer, Cham, https://doi.org/10.1007/978-3-030-69174-5_8

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249. DOI: 10.3322/caac.21660

Tekur, A., & Jain, P. (2018). A Study on Classification Algorithms for Predicting Colon Cancer using Gene Tissue Parameters. *International Journal of Pure and Applied Mathematics*, *119*(18), 2147-2166.

Tiwari, M., Bharuka, R., Shah, P., & Lokare, R. (2020). Breast cancer prediction using deep learning and machine learning techniques. *Available at SSRN 3558786*. Retrieved form https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3558786

Ummadi, J. R., Venkata Ramana Reddy, B., & Eswara Reddy, B. (2018). A Novel Statistical Feature Selection Measure for Decision Tree Models on Microarray Cancer Detection. In *Proceedings of International Conference on Computational Intelligence and Data Engineering* (pp. 229-245). Springer, Singapore. https://doi.org/10.1007/978-981-10-6319-0_20

Vamvakas, A., Tsivaka, D., Logothetis, A., Vassiou, K., & Tsougos, I. (2022). Breast Cancer Classification on Multiparametric MRI–Increased Performance of Boosting Ensemble Methods. *Technology in Cancer Research & Treatment*, *21*, 15330338221087828. https://doi.org/10.1177/15330338221087828

Vidyapith, B. (2020). Machine Learning Classifiers, Meta Classifiers Comparison And Analysis On Breast Cancer And Diabetes Datasets. *Advances and Applications in Mathematical Sciences*, *19*(10), 1017-1028.

Vinod, A., & Manju, B. R. (2020). Optimized Prediction Model to Diagnose Breast Cancer Risk and Its Management. In *Inventive Communication and Computational Technologies* (pp. 503-515). Springer, Singapore. https://doi.org/10.1007/978-981-15-0146-3_4.

Yadav, D. C., & Pal, S. (2019). Decision tree ensemble techniques to predict thyroid disease. *International Journal of Recent Technology and Engineering*, *8*(3), 8242-8246. DOI: 10.35940/ijrte.C6727.098319

Zhang, W., & Zhao, L. (2020). Online decision trees with fairness. arXiv preprint arXiv:2010.08146. https://doi.org/10.48550/arXiv.2010.08146

Zheng, J., Lin, D., Gao, Z., Wang, S., He, M., & Fan, J. (2020). Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. *IEEE Access*, *8*, 96946-96954. https://doi.org/10.1109/ACCESS.2020.2993536