www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

Advanced Filter Based Machine Learning Models on Clinical Databases for Outlier Detection

V. Devi Satya Sri[1]; Srikanth Vemuru[2]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, A.P, India.

[1]vdevisatyasri@gmail.com

[2]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, A.P, India.

Abstract

Feature selection approaches are used to improve the efficiency of the clinical databases in the machine learning classification. Since, most of the conventional feature selection and classification approaches are difficult to handle high dimensionality for pattern evaluation. Also these models are difficult to filter noise on different heterogeneous features. In this work, a hybrid data transformation and outlier detection methods are developed on the clinical databases to improve the classification accuracy. Experimental results show that the present model has better accuracy in evaluating the accuracy than the conventional models on clinical databases.

Key-words: Clinical Databases, Machine Learning, Classification.

1. Introduction

Medical data analysis enabled clinical results to be predicted by means of biological data such as biomarkers or disease patterns to diagnosis patients [1]. However, Medical data are highly dimensionally and sparingly characterized by the difficult use of traditional statistical methods and therefore classification and prediction-based machine-learning algorithms are used. A feature selection method based on correlation was also used to reduce the disease features to increase the algorithms accuracy [2]. However, there is no algorithm for machine learning which would consistently outperform one another, and the nature of this dataset appears to be a major influence on algorithm performance. However, by removing irrelevant genes the method for correlation-based

feature selection improved the prediction accuracy of all models [3]. A Medical data represents several thousand to several tens of thousands of gene/protein sequences. Each Medical data is scanned and transformed into normalized continuous data. The main issues of the Medical datasets are high dimensionality and imbalance nature. Traditional machine learning classifiers consider subset of features for classification and disease prediction with high true negative rate and error rates [4]. A biomedical dataset represents several thousand to several tens of thousands of gene/protein sequences. Each distributed medical data set data is scanned and transformed into normalized continuous data. The main issues of the distributed medical data are high dimensionality and imbalance nature. Traditional machine learning classifiers [5] consider subset of features for classification and disease prediction with high true negative rate and error rates [6]. To handle the uncertain data effectively using the information gain correctly, the well-known measure in information theory, i.e. entropy is used to find the impurity of the random samples [7]. Entropy assesses the impurity level in a group of samples [8]. The evaluation of information gain and gain ratio can be derived as follows:

$$Entropy = \sum_{i=0}^{n} -p_i \, log(p_i)$$

To determine the effectiveness of split metric in decision tree construction, [9] used information gain, which is partitioning the trained instances according to this metric measure. The gain measure of an attribute A, relative to a collection of instances I, is defined as,

$$Gain(I, A) = Entropy\ (I) - \sum_{i \in values(A)} \left|\frac{I_v}{I}\right| Entropy(I_v)$$

Where value (A) indicates the set of all distinct values for a metric A, and Iv is the subset of I for which the metric a has instance value v. Similarly, the metric selection measure of a discrete attribute A, relative to the given instances I, is given by split-info as:

$$SplitInfo\,(I\ ,A) = -\sum_{i=1}^{m} \left|\frac{I_i}{I}\right| \, log \left|\frac{I_i}{I}\right|$$

The gain ratio of a metric A is given as:

$$GainRatio = \frac{Gain\ (I, A)}{SplitInfo\ (I, A)}$$

Chi-Square is the common statistical test which measures the divergence from the expected distribution if the assumed feature occurrence is in fact independent of the class value. When the Chi-square Statistics is greater than the critical value defined by the degrees of freedom, then the function and class are deemed dependent. The T-test is another common statistical method used to compare a

sample mean or mean difference to the true mean or predicted mean difference. T – Test is yet another basic approach used in the study of feature patterns. The logarithm of expression levels was manipulated, requiring comparison calculation to the mean variance of both treatment control groups. The basic issue with T-test is that it involves repeated controlled trials with care, which are both repetitive and expensive. Relief-F is a feature selection technique that selects instances at random, adjusting feature-relevance weights depending on the nearest neighbour. By its merits Relief-F is one of feature selection's most effective strategies. Relief-F key concept is to calculate a score for each characteristic that calculates how well this feature distinguishes neighbouring samples in the original room. The nearest neighbour version searches for the closest example of the opposite class (nearest miss) in the original function space from the same class (nearest hit). The score is then the difference (or ratio) between the averages for all samples from the distance to the nearest miss the average distance on that function to the nearest hit in prediction [10]. In statistics, the selection of features also known as variable selection, is the method of selecting a subset of appropriate variables for statistical model construction.

Feature selection techniques are used under the central assumption that several redundant or inappropriate features are present in the data. Redundant apps are those that do not include any more detail than the currently selected apps. Methods of feature selection are also used in fields where there are many features and relatively few samples, such as data from DNA medical. Cancer studies are one of the areas where the study of Medical data is commonly used. Feature extraction is the process of extracting or calculating a value from each dataset sample. Features extracted by the techniques of static analysis are called static features, and features extracted by dynamic analysis are called dynamic features. The classifier is designed using a separate test set and evaluated from the training set. The data is split into N subsets with an almost equal size and distribution of classes. The classifier is made up of subsetsN-1, and the remaining subsets are used as the test set. Cross-validation can be repeated several times on random partitions. The average measurements calculated from different cross validation times are reliable estimates of a classifier's performance based on all the training data. First, they use some common selection of features to deal with it, but in the second step they use random projections to further reduce the dimensionality. This technique basically takes the matrix of features and multiplies it by a randomly selected matrix of projection that features vectors to other lower-dimensional space. This approach is equivalent to one hidden layer of a neural network that is assigned randomly but not learned by any algorithm. The paper shows that this approach leads to

significant savings in both learning time and assessment time. The main goal of automatic medical data feature extraction is to take source textual repositories, pre-process it and present the most significant information to the user in a condensed form and in a manner sensitive to user requirements. Feature extraction is the process of abstracting the main content from the different medical data sources and it has become an integral part of day to day activities in all domains like cloud, forums, social networking, and medical repositories. Automatic feature extraction fulfils certain goals by implementing feature extraction techniques at the user end to find relevant features of the large medical data set. Medical data features represent instances or phrases extracted from different sources without any subjective human intervention to find the essential patterns from large databases. Recently, Ensemble learning models [11] have become popular and widely accepted for high dimensional and imbalanced datasets. Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability. Feature selection is the process of selecting subset of features so that the original feature space is reduced using the selection measures. Feature selection measures can be categorized into three types: wrappers, filters, and embedded models [12] [13].

2. Related Works

Polsterl developed an advanced feature gene selection approach of adult ALL microarray data along with affinity propagation clustering scheme [14]. Microarray data analysis technique has become more popular and most implemented approach to identify diseases. It includes total ten thousand genes for input dimension. It is a severe computational issue in the process of data analysis. In this research paper, they introduced an affinity propagation clustering in case of feature gene selection of adult acute Lymphoblastic Leukaemia's microarray data. Identification of feature genes completely depends upon the total number of clusters in case of affinity propagation clustering. Affinity propagation clustering can be defined as an advanced clustering scheme that involves message interchange among various data points. Apart from this, it is also responsible for decreasing the dimension of every individual sample. It never requires prior knowledge about the total numbers of clusters. By analysing the outcomes of the above proposed approach, certain genes with AP clustering can provide appropriate learning in classification and prediction. Remeseiro proposed

partition-conditional ICA for Bayesian classification of microarray data [15]. Exact and proper classification of microarray data is a very crucial task in the field of medical decision making. By studying the previous research papers, we can say that class-conditional independent component analysis (CC-ICA) is efficient enough for enhancing the performance of naïve Bayes classifier in microarray data analysis. In certain cases, the microarray dataset contain few numbers of samples for several classes. In the above cases, some applications of CC-ICA appear to be infeasible. This research paper extends the traditional CC-ICA approach. The new extended and modified version of CC-ICA is known as partition-conditional independent component analysis (PC-ICA). As compared to the traditional approaches, PC-ICA provides better and most efficient feature extraction strategy. In other words, we can also mention here that, this above presented approach enhances the overall performance of Naïve Bayes classification of microarray data. ICA-NN: implemented an independent component analysis using neural action to transform a multi - variater distribution into mutual independent in statistical sense. Usually, these models are required to converge for desired actions to successfully estimate the independent medical data components [16]. ICA optimizes the objective function using imperialistic competition idea. Each member of this population is a vector of random numbers as biomedical node. The cost of the objective function will determine the biomedical action in relationship. ICA with neural action is apply successfully in classification, learning and optimization functions.

[17] Established that the selection of features involved specific redundant analyses. Thus, via redundant analysis, a new paradigm for efficient selection of features was developed. This system separated the analysis and redundant analysis applicable. A new feature selection algorithm was also introduced, and the best feature set was extracted against various learning algorithms. [18] Investigated the current ranking algorithms for the selection of functions. They found that the rating and classification varied markedly. Hence, an algorithm for selecting features was proposed. Two ranking models, Ranking SVM and RankNet, were used in this algorithm for extracting the best subset of features. [19] Proposed an algorithm for feature selection using Particle Swarm Optimization (PSO) and Vector Support Machines (SVMs). The PSO had been used to pick the best subset of apps. PSO's fitness function was tested using the SVM with the one-versus-rest approach. The algorithm has been tested with different classification problems.

[20] Identified a method of selecting features, namely consistent-based selection of features. It was a valuable indicator for various methods of selection of the devices. The proposed method thus

obtained a better result in accuracy than the wrapper approach and also obtained a higher reduction in functionality. [21] Proposed an algorithm for selection of features using the mining law of association and knowledge gain. The Apriori algorithm was used to find the attributes in question. Knowledge Gain was used in the dataset to delete the obsolete and redundant features.

The analysis of the algorithm showed that the classification accuracy had not improved. [22] Implemented a system of selection of features using rating criteria based on the filter. The proposed technique has been called TBFS (Threshold Dependent Feature Selection). Using the F-measure, the value of each attribute was normalized between 0 and 1 and the independent attribute was individually paired with the class attribute. This technique was useful in identifyingthe smaller subset of features and showed an increase in the accuracy of classification.

[23] Developed a subset selection algorithm for high-dimensional data based on the Clustering function. To separate the features into clusters, the graph-theoretic approach was used. The features which were closely related to the target class were chosen as the best subsets of features. They viewed every cluster as a single trait in this. Consequently, the dimensionality was reduced significantly. The algorithm was compared with various existing algorithms and the prediction accuracy and classification performance showed a minimum improvement.

[24] Used knowledge theory to examine the discriminatory selection feature algorithm which did not recognize the discriminating and continuous features of the dataset. They also proposed an algorithm for selection of features from the analysis. An Entropy breakpoint definition has been implemented in this algorithm. This algorithm has been tested using various real-world datasets. The experiment result stated that the algorithm had a high computational complexity with very low precision prediction. In Feature Selection, [25] proposed an algorithm for solving the optimisation problem. Using Greedy Search method and Greedy Search Loss from ranking method, this algorithm was used to find the best features. [26] Introduced a new algorithm called ReliefDisc. It functions on Discretization basis. Discretization partitions contains adjacent intervals in finite range. Rather than using random sampling to pick the instance, they suggested taking instance from each interval that reduces computational complexity and retains the consistency of features. There was also no need for user feedback to parameter the sample size. Experimental tests have shown that the current algorithm works better compared to the existing Relief algorithm. Relief-Disc achieved better than Relief according to them. The stochastic search approach is implemented in order to address the drawbacks of the conventional approaches, where some randomness is added in the search processand

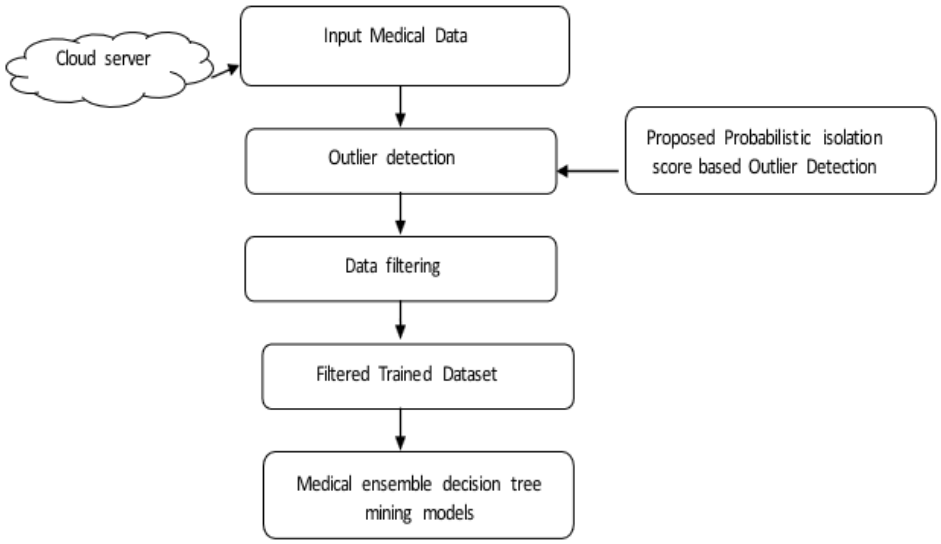
feature selection process is less sensitive to the specific dataset. The significance of the features is assessed in this process, their consistencies are calculated, and a pre-determined search algorithm is used along with a classifier. The consistency of a subset of features is calculated in terms of accuracy of the classification. The function subset contributing the highest accuracy of classification with a minimum number of features is considered to be the best. Different forms of Evolutionary Algorithms (EAs) are used to pick features. However, several algorithms for decision tree learning are degraded in their output performance because of inappropriate, unreliable or indefinite data are presented. Sometimes interdependent relationship among data are taken into consideration while some algorithms limitedly handles the attributes which have discrete values. Fast correlation-based filter method was used by Chen [27] in their proposed classification algorithm. The feature subsets are generated based on how these features correlate. The base classifier is the support vector machine which enables the algorithm to learn from the feature subsets presented to it. Voting is then done to determine the results. Sets of features are first created by subdividing the redundant features. This is in a bid to achieve diversity. This gives the classifier an opportunity to train from several subsets instead of one selected subset.[28] This method is more accurate than bagging and the other discussed methods. It has the capability of comfortably dealing with various types of features. Most predictive algorithms do not use the FCBF because it may lead to instability of the algorithm developed. Each particle in the ABC-PSO is a collection of n pheromone matrices where n is the number of nominal features in a biomedical data. Artificial Bee Colony (ABC) along with particle swarm optimization is one of the heuristic optimization approaches for action extraction. PSO optimizes [29-32] a problem by maintaining a population of particles (biomedical) and moving these particles (medical data) around in the search space. The action flow of the biomedical is guided by the best representative particles in the search space. Both the original ABC and PSO algorithms use sequential covering patterns to discover one-classification rule at a time.

3. Proposed Model

In the proposed framework, an advanced filter-based machine learning model is designed and implemented on the medical databases as shown in the figure 1. In this framework, different types of medical datasets are taken to find the outliers and data transformation process. After performing the data filtering operation and outlier detection models are applied to find the essential patterns for

decision making process. Finally, statistical measures are used to find the performance of the proposed model to the conventional models.

Figure 1 - Proposed Cloud based Medical Data Filtering Framework



Algorithm 1: Proposed Probabilistic isolation score based Outlier Detection

Input: Medical Training Data MD, Features space FS, Threshold Th.

Output: Outlier instances O, Normal instances N.

1. Read Input Medical dataset MD
2. Partition the input medical data into k subsets as KS[].
3. For each random subset k in KS[]
4. Do
5. Construct an Isolation tree to each partitioning subset of k in KS[]
6. For each instance in k[i]
7. Do
8. Calculate probability based class wise distribution score using the isolation path of the isolation tree as

$$m(s) = 2*(\log (s-1) + \lambda) - (2*(s-1) / s)$$

9. $$\mathrm{IProS}_i = 1 - \frac{\log (\mathrm{avgIPathlength})}{\mathrm{m (subsamplelen: s)}}$$
10. Mark the instances with highest average probability score as outliers

Received: 26.04.2021 – Accepted: 15.05.2021

11. If (Score>=Th)
12. Mark k(i) as outlier O
13. Else
14. Mark k(i) as normal N.
15. Done
16. Done
17. Apply data transformation and classification models on the non-outlier points for data learning andtesting.

Algorithm 1 illustrates the proposed isolation score-based outlier detection model for medical databases. In this algorithm, original data is partitioned into k subsets for probability score computation. Probability based class wise distribution score is computed to each instance in the kth subset of medical data. Here, isolation score is computed to each average path of the isolation tree. This score is used to check the outliers in the medical data for data transformation process.

Algorithm 2: High dimensional data normalization

Input: Non-outlier dataset ND, FS: Feature sets.

Output: Normalization data.

1. Read non-outlier medical data NMD with feature set FS.
2. For each feature in the FS
3. Do
4. Compute a non-linear transformation function to each feature values as given below as

$$\text{Nonlinear transform } (F[i]) = \eta = \left(\frac{1}{(1+\sqrt{2.\Pi.(1+e^{-|\log N.(F[i])|^2}})} \frac{1}{\cos(\sum \log(F[])} \right)$$

5. If (η >0.75)
6. Then
7. Normalize feature values within the range [η ,1]
8.
9. Else
10. Normalize feature values using min-max normalization with 0 and 1.

$$x' = \frac{x - \min\_(x)}{\max\_(x) - \min\_(x)} * (R2 - R1) + R1$$

11. End if
12. Done

Received: 26.04.2021 – Accepted: 15.05.2021

In this algorithm, a hybrid non-linear transformation algorithm is implemented on the non-outlier objects using the nonlinear transformation function. To each feature, a non-linear transformation function is computed in order to find the non-linear normalization range. If the computed normalization value is less than 0.75(threshold) then feature is transformed using the computed value. Otherwise, each value in the feature is normalized using the min-max normalization process.

Algorithm 3: Outlier detection based Classification model

Input: Proposed outlier detection points

Output: Classification patternsProcedure:

1. Read Outlier detection points D'.
2. Classifiers Cls[]={"H-Tree","C4.5","REP-TREE"}
3. To each classifier C[i] in Cls
4. Do
5. Apply C[i] on the D'.
6. Get TP rate of C[i] on the D'.
7. TP[]=TPRate(C[i]);
8. Done
9. Ensemble TPrate=Max{TP[]};
10. Construct decision tree of C[i] using the TPrate.

4. Experimental Results

Experimental results are simulated in java and NetBeans environment with Realtime cloud computing environment. In the proposed work, trauma cancer and tonsils dataset with large number of feature space. Proposed model is applied on the training medical datasets taken from cloud environment. Initially, these datasets are filtered by using the outlier detection algorithm and data transformation algorithm. Table 1 and table 2 represents the sample tonsils and trauma dataset for data filtering problem.

Table 1 - A Sample Tonsils Data

Gender	Throat Pain	Couch	f	Swelling	l	Outcome
0	1	0		1		No
0	1	1		1		Yes
0	0	0		0		No
0	1	0		1		No
1	0	0		0		No
0	1	0		0		No
0	0	1		0		No
0	0	1		1		Yes
1	1	0		0		No
0	1	1		0		No
1	1	0		1		No
0	0	0		1		No
1	1	0		1		No
1	0	1		0		No
1	0	1		1		Yes
0	1	1		0		No
1	1	0		1		No
0	0	1		1		Yes
0	1	0		0		No
0	1	1		0		No
0	1	1		0		No
1	1	1		1		Yes

Table 2 - Sample Trauma Dataset

ISS	NISS	PS14	Age	WBC [109/L]_T1	NEUT [109/L]_T1	LYMPH [109/L]_T1	MONO [109/L]_T1
11	11	97.2 149 3	41	26.16	21.95	1.6	2.58
9	22	99.5 979 5	22	8.28	4.21	3.53	0.33
14	17	99.1 46 2	47	9.24	4.52	3.73	0.45
4	12	99.8 511 4	38	11.54	6.37	4.18	0.65
24	34	87.3 260 3	78	17.43	12.48	3.46	1.1
10	11	98.7 174 3	32	7.74	3.42	3.57	0.53
16	29	93.8 455 6	45	19.75	8	9.96	1.18
9	18	99.7	19	7.54	4.77	2.15	0.51

		617 1					
20	29	99 956 7	27	11.6	4.89	5.75	0.73
16	16	99.5 835 3	26	12.9	8.27	3.83	0.66
29	41	99.1 480 8	24	7.89	4	2.94	0.55
25	34	96.5 935 1	29	15.43	10.16	4.24	0.77
24	34	98.8 731 5	41	15.06	7.37	6.79	0.65
9	27	99.7 617	40	5.7	3.17	1.67	0.64
		1					
34	34	76.8 653 4	20	20.06	16.6	2.67	0.57
5	9	91.2 387 7	45	0	0	0	0
25	57	98.8 211 7	24	10.28	5.94	3.23	0.91
45	75	50 389 9	75	13.81	10.12	2.54	1
50	66	19.2 79 4	64	11.49	4.32	6.54	0.51
45	66	69.2 196 7	19	11.86	4.03	7	0.69
20	29	76.7 798 3	71	9.53	5.24	3.69	0.45
29	75	52.1 785 5	75	8.84	6.15	1.94	0.57
42	66	23.9 732 3	54	10.04	4.57	4.68	0.57
50	66	98.2 63 4	20	13.94	6.77	5.87	1.01
29	57	56.5 596 8	22	13.53	8.34	4.3	0.8
29	41	51.9 769	20	20.61	15.06	4.26	1.06
29	41	61.4 512	56	14.09	10.2	2.91	0.9
38	75	88.4 383 9	25	12.82	6.46	5.63	0.46

38	43	74.160 2	66	11.39	4.55	5.76	0.5
29	57	34.341	49	4.75	3.14	1.24	0.34
9	22	85.1566 1	90	10.14	3.02	6.26	0.71
10	27	99.3484 6	45	9.15	3.83	3.84	1
9	22	99.6468 3	48	15.73	10	4.19	1.13
57	57	53.1169	25	7.2	3.67	2.59	0.8
9	27	99.7617 1	25	16.52	8.19	6.92	1.09
9	9	99.7617 1	21	11.38	6.13	4.1	0.89
41	48	55.1339	51	21.51	11.86	6.33	2.56
41	41	85234 3	76	8.25	5.7	2.11	0.36
19	27	96.4139 5	61	11.77	5.7	4.69	0.9
18	27	95.1964 4	43	19.19	15.23	3.06	0.68
9	12	98.8272 7	34	11.47	5.24	5.32	0.66

Table 3 Sample Tonsils and Trauma Outliers

Sample Tonsils Outliers

21.0,1.0,1.0,0.0,1.0,1.0,1.0,0.0,0.0,,No
59.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,1.0,,No
30.0,1.0,0.0,1.0,0.0,0.0,0.0,1.0,0.0,,No
56.0,1.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,,No
2.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,0.0,,Yes
36.0,1.0,1.0,1.0,0.0,0.0,1.0,0.0,1.0,,Yes
51.0,0.0,1.0,0.0,1.0,0.0,1.0,0.0,1.0,,No
11.0,0.0,1.0,1.0,0.0,1.0,0.0,1.0,1.0,,No
61.0,1.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,,No
45.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0,,No
14.0,0.0,1.0,0.0,1.0,1.0,1.0,0.0,0.0,,No
53.0,0.0,0.0,1.0,0.0,1.0,1.0,1.0,0.0,,Yes
35.0,1.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,,No
1.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,,No
29.0,0.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,,Yes

37.0,1.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,,No
58.0,1.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,,No
47.0,1.0,0.0,1.0,0.0,0.0,1.0,0.0,1.0,,Yes
9.0,0.0,1.0,1.0,1.0,1.0,0.0,1.0,0.0,,No
57.0,0.0,0.0,1.0,0.0,1.0,1.0,1.0,1.0,,Yes
20.0,1.0,0.0,1.0,0.0,1.0,1.0,0.0,0.0,,Yes
18.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,,Yes
11.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,,No
8.0,1.0,0.0,1.0,0.0,1.0,1.0,0.0,0.0,,Yes
27.0,0.0,0.0,1.0,1.0,0.0,0.0,1.0,1.0,,No
2.0,1.0,0.0,1.0,1.0,0.0,0.0,0.0,1.0,,No
45.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,,No
19.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,,No
12.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,,No
14.0,1.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,,No
20.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,1.0,,No
21.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,,No
10.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,1.0,,No
55.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,1.0,,No
15.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0,,No
52.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.0,,Yes
20.0,1.0,0.0,0.0,1.0,0.0,0.0,1.0,1.0,,No
41.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,,No
19.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,,No
33.0,0.0,1.0,1.0,0.0,1.0,0.0,0.0,0.0,,No
58.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,0.0,,Yes
35.0,0.0,1.0,0.0,1.0,1.0,0.0,1.0,1.0,,No
53.0,1.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,,No
30.0,0.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,,Yes
2.0,1.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,,No
43.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,1.0,1.0,,No
16.0,1.0,1.0,0.0,0.0,1.0,1.0,0.0,0.0,,No
11.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,,No
21.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,,No
39.0,1.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,1.0,,No
31.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,1.0,1.0,,No
25.0,0.0,0.0,1.0,0.0,0.0,1.0,0.0,0.0,,Yes
58.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,1.0,,No
56.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,1.0,1.0,,No
37.0,1.0,1.0,0.0,1.0,1.0,0.0,1.0,1.0,,No
15.0,1.0,1.0,1.0,1.0,0.0,0.0,1.0,1.0,,No
60.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,1.0,,No
50.0,0.0,1.0,1.0,0.0,0.0,0.0,1.0,1.0,,No
54.0,1.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,,No
62.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,1.0,,No
60.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,,No
26.0,0.0,1.0,0.0,1.0,1.0,1.0,0.0,1.0,,No

16.0,0.0,1.0,0.0,1.0,0.0,0.0,1.0,0.0,,No
35.0,1.0,1.0,0.0,0.0,1.0,0.0,1.0,0.0,,No
54.0,0.0,0.0,0.0,1.0,1.0,1.0,0.0,0.0,,No
46.0,1.0,0.0,1.0,0.0,0.0,1.0,0.0,0.0,,Yes
21.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,,No
35.0,0.0,1.0,0.0,1.0,1.0,1.0,1.0,1.0,,No
44.0,1.0,1.0,1.0,0.0,0.0,1.0,0.0,1.0,,Yes
4.0,1.0,1.0,0.0,0.0,1.0,0.0,0.0,1.0,,No
44.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,,No
24.0,1.0,0.0,1.0,1.0,0.0,0.0,1.0,1.0,,No
29.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,,No
11.0,0.0,1.0,0.0,0.0,1.0,1.0,1.0,0.0,,No
51.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,1.0,,No
59.0,1.0,0.0,1.0,0.0,1.0,0.0,1.0,0.0,,No
41.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,,Yes
4.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,,No
32.0,1.0,1.0,0.0,1.0,1.0,1.0,0.0,0.0,,No
50.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,No
42.0,0.0,0.0,1.0,1.0,0.0,1.0,1.0,0.0,,Yes
61.0,0.0,0.0,0.0,1.0,1.0,1.0,1.0,0.0,,No
40.0,1.0,1.0,0.0,1.0,1.0,0.0,0.0,0.0,,No
45.0,0.0,1.0,0.0,0.0,1.0,1.0,0.0,0.0,,No
47.0,1.0,0.0,1.0,1.0,1.0,1.0,1.0,0.0,,Yes
9.0,0.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,,Yes
56.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,,No
13.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,0.0,,Yes
30.0,0.0,0.0,1.0,1.0,1.0,0.0,0.0,1.0,,No

Sample Trauma outliers

0.0 2.0 55.0 14.0 26.0 28.0 28.0 14.0 8.0 24.0 34.0 16.0 8.0 5.0 7.0 19.0 48.0 40.0 15.0 44.0 53.0
11.0 30.0 55.0 0.0 0.0 8.0 14.0 16.0 1.0 13.0 37.0 15.0 21.0 8.0 32.0 27.0 20.0 27.0 7.0 3.0 20.0 5.0
12.0 0.0 0.0 0.0 0.0 0.0 21.0 26.0 19.0 29.0 35.0 28.0 20.0 14.0 21.0 33.0 29.0 46.0 46.0 26.0 45.0
23.0 33.0 33.0 4.0 55.0 56.0 48.0 39.0 23.0 42.0 18.0 8.0 11.0 15.0 28.0 46.0 33.0 8.0 6.0 13.0 31.0
1.0 26.0 36.0 0.0 0.0 25.0 52.0 14.0 1.0 20.0 31.0 35.0 36.0 11.0 35.0 30.0 42.0 15.0 4.0 2.0 14.0
14.0 24.0 0.0 0.0 0.0 0.0 0.0 39.0 38.0 13.0 39.0 39.0 38.0 36.0 15.0 27.0 31.0 35.0 36.0 42.0 23.0
46.0 8.0 29.0 34.0 18.0 5.0 6.0 9.0 16.0 3.0 19.0 23.0 43.0 5.0 7.0 9.0 49.0 6.0 5.0 49.0 5.0 24.0
13.0 11.0 52.0 0.0 0.0 3.0 3.0 11.0 0.0 16.0 34.0 18.0 24.0 18.0 13.0 46.0 36.0 19.0 5.0 1.0 13.0 2.0
11.0 0.0 0.0 0.0 0.0 0.0 11.0 12.0 3.0 10.0 19.0 10.0 14.0 8.0 14.0 20.0 14.0 46.0 46.0 43.0 44.0
38.0 32.0 35.0 41.0 N
2.0 7.0 54.0 6.0 41.0 34.0 52.0 30.0 14.0 16.0 40.0 25.0 9.0 0.0 0.0 5.0 11.0 23.0 40.0 36.0 28.0
37.0 21.0 24.0 28.0 9.0 23.0 34.0 6.0 0.0 34.0 17.0 38.0 17.0 7.0 15.0 36.0 36.0 30.0 9.0 1.0 21.0
6.0 32.0 35.0 37.0 37.0 21.0 29.0 45.0 47.0 27.0 46.0 40.0 36.0 41.0 35.0 45.0 19.0 34.0 34.0 19.0
29.0 21.0 40.0 1.0 29.0 1.0 47.0 44.0 52.0 44.0 17.0 15.0 37.0 28.0 8.0 0.0 0.0 12.0 26.0 16.0 34.0
19.0 12.0 32.0 29.0 15.0 15.0 33.0 20.0 37.0 1.0 0.0 46.0 5.0 43.0 6.0 11.0 31.0 21.0 37.0 21.0 6.0
2.0 13.0 12.0 17.0 37.0 35.0 37.0 12.0 22.0 46.0 45.0 9.0 46.0 45.0 26.0 46.0 25.0 43.0 12.0 22.0
27.0 26.0 36.0 26.0 33.0 7.0 27.0 19.0 21.0 22.0 27.0 10.0 4.0 24.0 31.0 14.0 5.0 0.0 0.0 23.0 11.0

32.0 20.0 22.0 10.0 20.0 37.0 15.0 29.0 22.0 12.0 15.0 3.0 0.0 48.0 5.0 14.0 2.0 20.0 25.0 21.0 13.0 18.0 4.0 1.0 15.0 10.0 20.0 32.0 11.0 8.0 8.0 16.0 32.0 31.0 9.0 36.0 26.0 19.0 35.0 23.0 38.0 7.0 18.0 30.0 30.0 15.0 14.0 10.0 1.0 16.0 10.0 N

18.0 10.0 26.0 22.0 37.0 25.0 54.0 35.0 9.0 9.0 50.0 33.0 7.0 10.0 32.0 34.0 12.0 46.0 32.0 11.0 14.0 31.0 27.0 5.0 9.0 46.0 23.0 30.0 13.0 1.0 14.0 36.0 39.0 32.0 8.0 23.0 45.0 41.0 18.0 5.0 3.0 28.0 5.0 30.0 27.0 26.0 26.0 7.0 41.0 48.0 48.0 34.0 3.0 2.0 49.0 48.0 47.0 5.0 3.0 46.0 14.0 7.0 4.0 4.0 11.0 14.0 2.0 33.0 38.0 42.0 30.0 28.0 11.0 38.0 17.0 13.0 5.0 17.0 32.0 16.0 14.0 43.0 11.0 5.0 21.0 25.0 44.0 1.0 31.0 15.0 16.0 26.0 5.0 1.0 32.0 19.0 24.0 19.0 7.0 8.0 52.0 45.0 4.0 2.0 2.0 6.0 13.0 19.0 24.0 29.0 30.0 3.0 35.0 30.0 25.0 10.0 2.0 3.0 45.0 38.0 35.0 6.0 2.0 37.0 35.0 23.0 13.0 22.0 1.0 8.0 4.0 33.0 54.0 55.0 47.0 25.0 10.0 33.0 27.0 8.0 5.0 29.0 37.0 5.0 12.0 6.0 56.0 1.0 4.0 6.0 3.0 5.0 31.0 5.0 22.0 43.0 22.0 1.0 2.0 52.0 16.0 36.0 8.0 8.0 40.0 40.0 13.0 2.0 1.0 16.0 7.0 10.0 46.0 22.0 20.0 5.0 38.0 42.0 44.0 21.0 11.0 14.0 42.0 45.0 41.0 16.0 3.0 38.0 42.0 47.0 38.0 17.0 43.0 31.0 33.0 28.0 Y

7.0 8.0 27.0 19.0 45.0 33.0 57.0 34.0 23.0 7.0 52.0 20.0 14.0 4.0 9.0 7.0 26.0 45.0 41.0 15.0 39.0 24.0 14.0 17.0 27.0 10.0 25.0 35.0 12.0 1.0 18.0 32.0 41.0 28.0 9.0 34.0 4.0 18.0 15.0 4.0 3.0 34.0 6.0 33.0 45.0 41.0 41.0 28.0 22.0 49.0 49.0 23.0 49.0 46.0 47.0 45.0 27.0 29.0 34.0 48.0 23.0 11.0 6.0 9.0 11.0 20.0 4.0 44.0 28.0 27.0 50.0 20.0 7.0 14.0 42.0 12.0 3.0 12.0 7.0 44.0 35.0 45.0 36.0 30.0 44.0 52.0 41.0 12.0 32.0 24.0 24.0 28.0 4.0 1.0 41.0 10.0 18.0 7.0 11.0 38.0 9.0 8.0 3.0 2.0 2.0 14.0 14.0 22.0 32.0 17.0 17.0 11.0 11.0 48.0 47.0 8.0 47.0 48.0 36.0 31.0 10.0 35.0 21.0 30.0 23.0 6.0 2.0 5.0 7.0 16.0 5.0 9.0 33.0 31.0 40.0 17.0 3.0 17.0 38.0 18.0 3.0 15.0 22.0 26.0 48.0 38.0 16.0 23.0 53.0 36.0 47.0 11.0 14.0 35.0 25.0 28.0 10.0 1.0 45.0 8.0 22.0 10.0 24.0 28.0 19.0 18.0 6.0 2.0 1.0 6.0 16.0 18.0 28.0 27.0 23.0 14.0 11.0 48.0 48.0 12.0 47.0 40.0 37.0 25.0 17.0 24.0 16.0 29.0 9.0 15.0 4.0 7.0 3.0 14.0 4.0 17.0 N

14.0 7.0 15.0 3.0 53.0 47.0 42.0 38.0 33.0 40.0 18.0 20.0 21.0 10.0 38.0 22.0 57.0 22.0 17.0 28.0 30.0 35.0 46.0 19.0 0.0 0.0 42.0 50.0 4.0 0.0 40.0 10.0 44.0 15.0 8.0 11.0 37.0 39.0 10.0 2.0 3.0 17.0 4.0 7.0 42.0 22.0 22.0 12.0 8.0 36.0 40.0 28.0 40.0 23.0 39.0 26.0 6.0 44.0 10.0 22.0 44.0 47.0 40.0 0.0 43.0 35.0 34.0 12.0 33.0 30.0 36.0 46.0 4.0 12.0 30.0 39.0 2.0 7.0 11.0 7.0 52.0 30.0 38.0 45.0 48.0 9.0 9.0 49.0 0.0 0.0 27.0 32.0 6.0 0.0 27.0 24.0 44.0 29.0 11.0 26.0 25.0 43.0 27.0 8.0 2.0 29.0 5.0 7.0 42.0 8.0 8.0 5.0 9.0 22.0 19.0 8.0 21.0 17.0 28.0 27.0 5.0 36.0 7.0 21.0 25.0 37.0 37.0 0.0 27.0 31.0 35.0 20.0 10.0 11.0 28.0 8.0 2.0 14.0 41.0 19.0 3.0 6.0 19.0 19.0 47.0 34.0 19.0 26.0 30.0 41.0 48.0 28.0 0.0 0.0 7.0 7.0 7.0 0.0 31.0 23.0 11.0 16.0 21.0 32.0 14.0 9.0 22.0 7.0 1.0 25.0 13.0 13.0 42.0 3.0 3.0 1.0 9.0 30.0 29.0 12.0 38.0 16.0 23.0 29.0 4.0 41.0 4.0 8.0 41.0 34.0 36.0 0.0 29.0 33.0 36.0 31.0 Y

2.0 0.0 54.0 3.0 22.0 23.0 26.0 24.0 2.0 22.0 33.0 32.0 2.0 10.0 17.0 31.0 8.0 58.0 10.0 41.0 45.0 30.0 50.0 39.0 10.0 42.0 12.0 22.0 15.0 1.0 7.0 43.0 24.0 33.0 7.0 31.0 15.0 24.0 24.0 6.0 3.0 32.0 4.0 16.0 19.0 23.0 23.0 17.0 6.0 23.0 29.0 28.0 38.0 13.0 24.0 22.0 21.0 39.0 6.0 5.0 12.0 14.0 29.0 39.0 19.0 4.0 16.0 15.0 40.0 36.0 47.0 42.0 5.0 13.0 36.0 33.0 2.0 14.0 24.0 37.0 12.0 58.0 29.0 33.0 37.0 27.0 32.0 44.0 24.0 28.0 1.0 2.0 16.0 0.0 2.0 49.0 33.0 44.0 11.0 39.0 14.0 29.0 26.0 8.0 2.0 34.0 7.0 24.0 29.0 23.0 25.0 8.0 8.0 42.0 40.0 21.0 45.0 29.0 39.0 43.0 17.0 42.0 13.0 19.0 14.0 4.0 21.0 37.0 17.0 1.0 21.0 4.0 6.0 7.0 22.0 10.0 2.0 11.0 36.0 25.0 3.0 17.0 36.0 47.0 31.0 60.0 48.0 7.0 39.0 45.0 23.0 55.0 23.0 36.0 13.0 11.0 10.0 0.0 15.0 39.0 11.0 19.0 25.0 35.0 9.0 7.0 29.0 11.0 1.0 37.0 9.0 21.0 14.0 4.0 4.0 4.0 1.0 29.0 34.0 18.0 41.0 11.0 21.0 30.0 12.0 40.0 7.0 4.0 40.0 4.0 26.0 41.0 8.0 7.0 22.0 12.0 N

Table 3 illustrates the outliers detected using the proposed outlier detection approach on the tonsils dataset and trauma dataset.

Table 4 - Comparative Analysis of Number of Outliers Detected in the Tonsils Dataset

Sample	I Tree	Gaussian	Proposed Outlier Model
#1	46	58	84
#2	59	57	81
#3	30	63	92
#4	51	64	80
#5	60	55	89
#6	30	64	87
#7	54	51	91
#8	37	65	80
#9	62	52	91

Table 5 illustrates the comparative analysis of different outlier detection approaches to the proposed outlier detection approach on trauma dataset. From the table, the conventional approaches have a smaller number of outliers than the proposed approach on trauma dataset.

Figure 2 - Performance Results of Advanced Boosting Classifier Runtime to the Conventional Approaches on the Tonsils Dataset

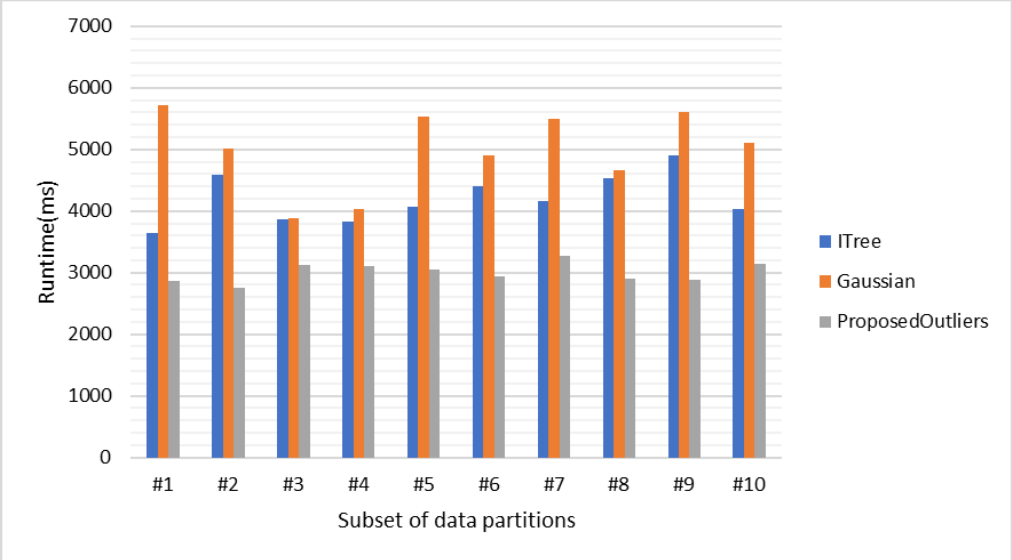


Figure 2 illustrates the comparative results of advanced boosting classifier to the conventional approaches on the tonsil dataset. From the figure, it is observed that the present model has better runtime than the conventional approaches on the tonsil dataset.

Figure 3 - Performance Results of Advanced Boosting Classifier Runtime to the Conventional Approaches on the Trauma Dataset

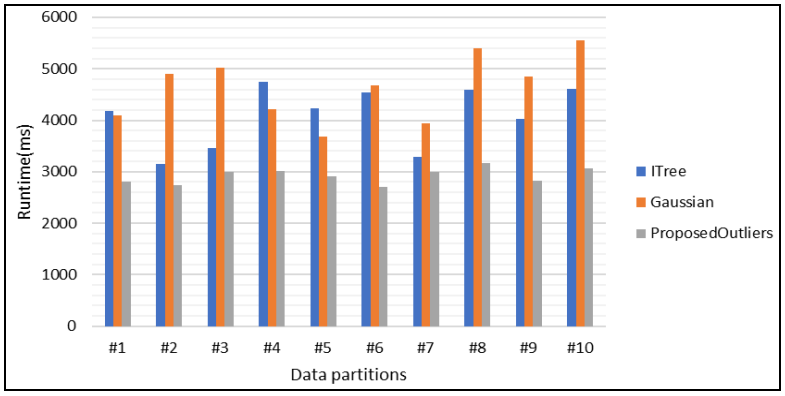


Figure 3 illustrates the comparative results of advanced boosting classifier to the conventional approaches on the trauma dataset. From the figure, it is observed that the present model has better runtime than the conventional approaches on the trauma dataset.

Decision Patterns Generated by Using the Ensemble Learning Decision Trees

WAIS_PSI_Composite_6mo = 99.203279 AND
Rotterdam = 2 AND
GOSE_OverallScore3M = 8: 1 (14.0/4.0)

WAIS_PSI_Composite_6mo = 99.203279 AND
Rotterdam = 2 AND
GOSE_OverallScore3M = 7: 1 (11.0/4.0)

WAIS_PSI_Composite_6mo = 99.203279 AND
Rotterdam = 2 AND
GOSE_OverallScore3M = 6 ANDFACIALFX = 0: 1
(6.0/2.0)

WAIS_PSI_Composite_6mo = 99.203279 AND
GOSE_OverallScore3M = 6.217105 AND
SDH_FINAL = 0 AND
Marshall = 1 AND MR_result = 0.5625 AND
rs4680 =3: 2 (10.0/3.0)

WAIS_PSI_Composite_6mo = 99.203279 AND

GOSE_OverallScore3M = 6.217105 AND

SDH_FINAL = 0 AND

FACIALFX = 0 AND

Marshall = 1 ANDrs4680 = 2 AND

MR_result = 0.5625: 1 (6.0/3.0)

WAIS_PSI_Composite_6mo = 99.203279 AND

GOSE_OverallScore3M = 6.217105 AND

Marshall = 1 AND

rs4680 = 3: 1 (4.0/1.0)

WAIS_PSI_Composite_6mo = 99.203279 ANDGOSE_OverallScore3M = 6.217105 AND

Marshall = 1 AND

FACIALFX = 0 AND

MR_result = 0.5625: 1 (5.0/1.0)

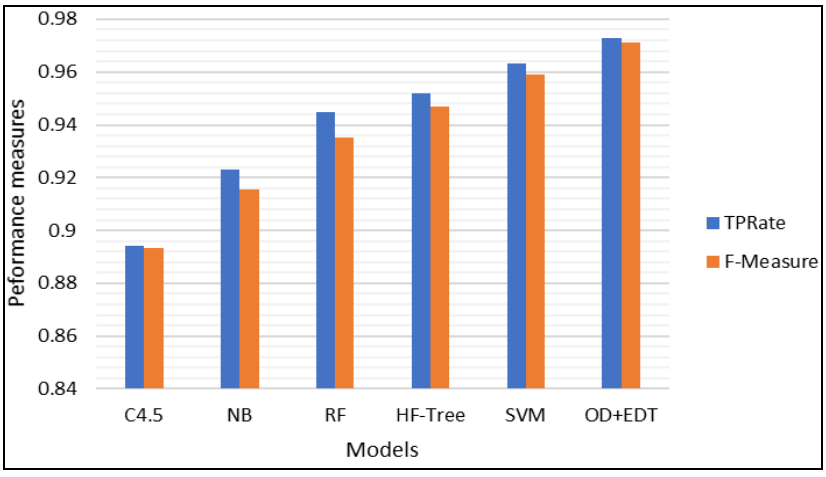
WAIS_PSI_Composite_6mo = 99.203279 AND

GOSE_OverallScore3M = 6.217105 AND

FACIALFX = 0 AND

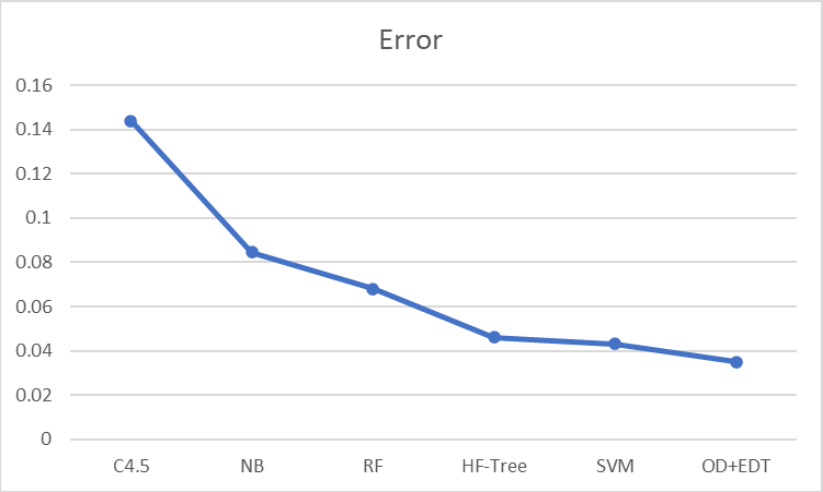
Marshall = 2: 1 (8.0/4.0)

Figure 4 - Statistical Analysis of Proposed Outlier Detection based Ensemble Decision Tree Model to the Conventional Models



Received: 26.04.2021 – Accepted: 15.05.2021

Figure 5 - Error Rate of Proposed Outlier Detection based Ensemble Decision Tree Model to the Conventional Models



5. Conclusion

In this work, an advanced machine learning approaches are implemented on the medical databases for better decision making. Since, most of the medical databases contain high dimensionality and large data size, it is difficult to find an essential key feature using traditional feature sub-set selection approaches. Also, conventional medical data filtering techniques fail to find the essential outliers due to large data size and feature space. In this work, a hybrid outlier detection and data transformation approaches are implemented to remove the noise in the medical databases. Experimental results proved that the present model has high outlier detection rate than the traditional approaches on different medical databases.

References

Abdel-Aal, R.E. (2005). GMDH-based feature ranking and selection for improved classification of medical data. *Journal of Biomedical Informatics, 38*(6), 456-468. https://doi.org/10.1016/j.jbi.2005.03.003

Alirezanejad, M., Enayatifar, R., Motameni, H., & Nematzadeh, H. (2020). Heuristic filter feature selection methods for medical datasets. *Genomics, 112*(2), 1173-1181., Mar. 2020, doi: 10.1016/j.ygeno.2019.07.002.

Chatterjee, R., Maitra, T., Islam, S.H., Hassan, M.M., Alamri, A., & Fortino, G. (2019). A novel machine learning based feature selection for motor imagery EEG signal classification in Internet of medical things environment. *Future Generation Computer Systems, 98,* 419-434. doi: 10.1016/j.future.2019.01.048

Danilov, V. V., Skirnevskiy, I.P., Manakov, R. A., Gerget, O.M., & Melgani, F. (2020). Feature selection algorithm based on PDF/PMF area difference. *Biomedical Signal Processing and Control, 57,* 101681. doi: 10.1016/j.bspc.2019.101681.

Du, G., Zhang, J., Luo, Z., Ma, F., Ma, L., & Li, S. (2020). Joint imbalanced classification and feature selection for hospital readmissions. *Knowledge-Based Systems, 200,* 106020. Jul. 2020, doi: 10.1016/j.knosys.2020.106020.

Haider, F., Pollak, S., Albert, P., & Luz, S. (2021). Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech & Language, 65,* 101119. Jan. 2021, doi: 10.1016/j.csl.2020.101119.

Hu, J., Li, Y., Gao, W., & Zhang, P. (2020). Robust multi-label feature selection with dual-graph regularization. *Knowledge-Based Systems, 203,* 106126. Sep. 2020, doi: 10.1016/j.knosys.2020.106126.

Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., & Li, J. (2020). Sample imbalance disease classification model based on association rule feature selection. Pattern Recognition Letters, 133, 280-286. doi: 10.1016/j.patrec.2020.03.016.

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H., Xia, Y., Mamlin, B., & Grannis, S. J. (2016). Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. Journal of biomedical informatics, 60, 145-152. doi: 10.1016/j.jbi.2016.01.008.

Lee, I. G., Zhang, Q., Yoon, S. W., & Won, D. (2020). A mixed integer linear programming support vector machine for cost-effective feature selection. Knowledge-Based Systems, 203, 106145. doi: 10.1016/j.knosys.2020.106145

Lee, J., Jeong, J. Y., & Jun, C.H. (2020). Markov blanket-based universal feature selection for classification and regression of mixed-type data. *Expert Systems with Applications, 158,* 113398. doi: 10.1016/j.eswa.2020.113398

Liu, Q., Gu, Q., & Wu, Z. (2017). Feature selection method based on support vector machine and shape analysis for high-throughput medical data. *Computers in biology and medicine, 91,* 103-111. Dec. 2017, doi: 10.1016/j.compbiomed.2017.10.008

Nguyen, B. H., Xue, B., & Zhang, M. (2020). A survey on swarm intelligence approaches to feature selection in data mining. *Swarm and Evolutionary Computation, 54,* 100663. doi: 10.1016/j.swevo.2020.100663.

Pölsterl, S., Conjeti, S., Navab, N., & Katouzian, A. (2016). Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection. *Artificial intelligence in medicine, 72,* 1-11. doi: 10.1016/j.artmed.2016.07.004.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in biology and medicine, 112,* 103375. doi: 10.1016/j.compbiomed.2019.103375.

Shah, S. M. S., Shah, F. A., Hussain, S. A., & Batool, S. (2020). Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods. *Computers & Electrical Engineering, 84,* 106628. doi: 10.1016/j.compeleceng.2020.106628.

Shu, W., Qian, W., & Xie, Y. (2020). Incremental feature selection for dynamic hybrid data using neighborhood rough set. *Knowledge-Based Systems, 194,* 105516. doi: 10.1016/j.knosys.2020.105516

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Computer Science, 167,* 1970-1980. doi: 10.1016/j.procs.2020.03.226

Sun, L., Yin, T., Ding, W., Qian, Y., & Xu, J. (2020). Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems. Information Sciences, 537, 401-424. Oct. 2020, doi: 10.1016/j.ins.2020.05.102.

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Classification of flower species by using features extracted from the intersection of feature selection methods in convolutional neural network models. *Measurement, 158,* 107703. Jul. 2020, doi: 10.1016/j.measurement.2020.107703.

Tsai, C.F., & Chen, Y.C. (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences, 505,* 282-293. Dec. 2019, doi: 10.1016/j.ins.2019.07.091

Tuba, E., Strumberger, I., Bezdan, T., Bacanin, N., & Tuba, M. (2019). Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. *Procedia Computer Science, 162,* 307-315. doi: 10.1016/j.procs.2019.11.289.

Uçar, M.K. (2020). Classification performance-based feature selection algorithm for machine learning: P-score. *IRBM, 41*(4), 229-239. Feb. 2020, doi: 10.1016/j.irbm.2020.01.006.

Wiesław, P. (2019). Tree-based generational feature selection in medical applications. *Procedia Computer Science, 159,* 2172-2178. Jan. 2019, doi: 10.1016/j.procs.2019.09.391.

Xie, J., Li, Y., Wang, N., Xin, L., Fang, Y., & Liu, J. (2020). Feature selection and syndrome classification for rheumatoid arthritis patients with Traditional Chinese Medicine treatment. European *Journal of Integrative Medicine, 34,* 101059. Feb. 2020, doi: 10.1016/j.eujim.2020.101059.

Castro-Zunti, R., Park, E. H., Choi, Y., Jin, G. Y., & Ko, S. B. (2020). Early detection of ankylosing spondylitis using texture features and statistical machine learning, and deep learning, with some patient age analysis. *Computerized Medical Imaging and Graphics, 82,* 101718. doi: 10.1016/j.compmedimag.2020.101718

Chen, Y., Zhu, G., Liu, D., Liu, Y., Yuan, T., Zhang, X., ... & Zhang, J. (2020). The morphology of thalamic subnuclei in Parkinson's disease and the effects of machine learning on disease diagnosis and clinical evaluation. *Journal of the neurological sciences, 411,* 116721. doi: 10.1016/j.jns.2020.116721

Sri[1], V. D. S., & Vemuru, S. (2019). Survey on Data Security Issues related to Multi-user Environment in Cloud Computing. *Journal of Critical Reviews, 7*(4), 2020.