IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A HYBRID APPROACH FOR INTENSIFYING THE CLASSIFICATION ACCURACY IN HEALTHCARE DATA

[1]Manikandan J, [2]K. Palanivel

[1]M.Tech(CSE), [2]Systems Analyst
[1]Department of Computer Science,
[1]Pondicherry University, Pondicherry, India

Abstract: Data mining and machine learning plays an extensively crucial role in the application of medical diagnosis. Consequently, several forms of analysis are ongoing to better predict the diseases and improved the quality of diagnosis. The existing analysis suffers with high dimensionality and imbalanced class distribution problem. Due to this issue, various classification algorithms prejudge over the majority classes, while ignoring the minority classes. This leads to the misprediction while predicting some rarely possible diseases. This research paper aims to increase the classification accuracy of various classifier over healthcare data. This research proposed a hybrid approach of both Principle Component Analysis (PCA) and Synthetic Minority Oversampling Technique (SMOTE) to reduce the high dimensional imbalanced class distribution problem. Then the resultant dataset is applied to the various classifier, and compared based on the evaluation metrics.

Index Terms **- Data mining, machine learning, high dimensionality problem, imbalanced class distribution problem, Principle Component Analysis, SMOTE.**

I. INTRODUCTION

In the information age, the data are plentiful. The demands of information management, research, and analysis have become more and more important. Todays, the amount of data generated and gathered has been increasing rapidly. This explosive growth in data need for new techniques and tools that can intelligently and automatically analyze relevant data. The traditional methods were too time-consuming to cope with the massive quantities of date. As a result, data mining has become a very important research topic to enhance the value of existing information resources [41]. Currently data mining is an emerging trends and it is widely used in almost all the application like medical diagnosis, share market analysis, fraud detection, banking, finance, spam Email filtering to discern the sentiment, educational institute and others. .

According to Holsheimer M, Siebes A, 1994 [47], data mining can be defined as the non-trivial process of extracting implicit, a priori unknown useful information (such as knowledge rules, constraints, and regularities) from the stored data. Data mining techniques can be implemented rapidly on existing platforms to enhance the value of existing information resources, and can be integrated with systems as they are brought on-line. Data mining is a process of extracting or discovering a knowledge from huge volume of datasets and it is currently analyzing and exploring a large volume of information to glean meaningful trends and patterns [1]. It is different from the other technology because it follow broader concepts, with various steps to it. In data mining, the data is being pre-processed, normalized, redundancy of data are rectified and noise will be minimized and then the data mining technique such as association rule mining, clustering, classification, etc., will be applied. Finally, the result of the applied mining technique will be evaluated and analyzed.

In data mining, classification is a task of data analysis. It is termed as the process of building a classifier model that demonstrates and distinguishes the various concepts and classes of the data. The classification process is mainly consist of two phases called training set and testing set [18]. Initially the dataset is divided into some specific ratios according to its size for train and test purposes. In training phase, to construct the classifier model and to give learning to the model using training dataset. Whereas in testing phase, classification model is used to predict the corresponding class label while randomly, give some test data. It is a type of supervised learning mechanism.

Fuzzy data mining methods can mean *data mining methods* that are fuzzy methods as well. On the other hand, it can also mean approaches to analyze fuzzy data. Fuzzy methods conceptions are different. Fuzzy data mean imprecise, vague, uncertain, ambiguous, inconsistent and/or incomplete data. Therefore, the source of uncertainty is the data themselves. It is very important to develop methods that are able to handle this kind of data because data from several information sources might be fuzzy. Rudolf Kruse [42] stated that data mining is essentially a circular processthat can trigger a re-execution of the data preparation and model generation steps. In this process, fuzzy set methods can profitably be applied in several phases - *Business understanding and Data understanding phase, data preparation phase, modeling phase and finally evaluation phase.*

According to Zadeh LA, 1965 [46], Fuzzy logic works with reasoning rules which is approximate and intuitive. The fuzzy logic allows us to define values without specifying a precise value. Fuzzy sets are a generalization of traditional sets. The μVLpH (X) = 0 and μVLpH (X) = 1 cases which would correspond to conventional sets, are just special cases of fuzzy sets. The use of fuzzy sets defined by means of membership functions in logic expressions is called *Fuzzy Logic*. Hence, it is possible to write a set of rules representing the relation between input and output variables. These rules present the format 'if-then', and are made up of an antecedent and a consequent. A fuzzy rule 'if-then' is a structure for representing imprecise knowledge. The following illustrate briefly these logic interferences:

a. If x is A then y is C (1)
b. If x is A and z is B then y is C (2)

The process of extracting knowledge from a database is called Knowledge Discovery in Databases (or KDD). KDD has evolved from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence, computational intelligence, etc. This process is made up of several stages ranging from data preparation to achievement of results. The additional steps in the KDD process are essential to ensure that useful knowledge is derived form the data. They are data selection, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the results. In KDD, one of the stages is called data mining (DM).

1.1 Existing Problem & Challenges

As the basis for data mining, healthcare data deliver enormous benefits. Data mining is becoming increasingly prevalent in healthcare. Applications for data mining will greatly benefit all interested parties in the healthcare sector. Data mining will benefit healthcare sector as doctors recognize appropriate therapies and best practices, and patients access more and more accessible healthcare services. Classification algorithms are widely used in healthcare sector. It usually construct a classifier model to classify the dataset based on its class label. However, it classifies the dataset in an efficient way it would always incapable to produce the reliable model on the enormous amount of datasets. The huge volume of dataset holds numerous redundant and irrelevant features that misinform the classifiers. Furthermore, an enormous amount of dataset have an imbalanced class distribution, which leads prejudice over the majority class in the classification process. Therefore, the classifier misclassify with the majority classes and ignore the minority classes. So the prediction of medical diagnosis will be highly affected while predicting very rarely possible diseases. In this study, it will concentrate on imbalanced dataset and to increases the classification accuracy. Some widely used classification algorithms are Tree based, Rule based, Lazy learners and neural networks. In this paper, it is proposed to use some of the algorithm in the experimentation.

1.2 Proposed Solution

Due to the advancement of technologies and heterogeneous platforms huge amount of data is generated which leads to mislabeled records, incorrect data, missing values and noise. In the existing system, they invented so many Resampling technique such as random under sampling, random over sampling, cluster based over sampling and SMOTE are used in order to increase the classification accuracy. However, they only focus on balance the class distribution and it not work with high dimensionality data problem and it will also increasing the noise factor. To address these issues, we proposed a hybrid approach. Initially PCA and Feature selection techniques are used as to reduce the high dimensionality of the dataset, which can lead to the dataset as dimensions free. This work eliminates the imbalance issues by using SMOTE algorithm. It is an oversampling algorithm and it balances the dataset by adding synthetic instances by calculating the Distances between the samples of training dataset and the samples of minority class.

The rest of paper is organized as follows: Section 2 described existing literature and related work regarding the paper. Section 3 described about the proposed algorithm, and their methodologies is clearly explained in corresponding subsections. Experimental evaluation of the proposed method, description about the dataset and their respective results is clearly described in section 4. Finally, the paper is concluded in section 5.

II. LITERATURE REVIEW AND RELATED WORKS

A comprehensive survey of data mining summarized the requirements and challenges of data mining, which included handling different types of data, efficiency and scalability of data mining algorithms, usefulness, certainty, and expressiveness of data mining results, expression of various kinds of data mining results, interactive mining knowledge at multiple abstraction levels, mining information from different sources of data, and protection of privacy and data security. Some of these requirements may carry conflicting goals. Therefore, different techniques are used to address some of these problems.

Researchers have performed different experiments on medical datasets by exploiting feature selection and multiple classifiers techniques. The enormous growth of data size and the number of existing databases exceeds the human capacity to analyze the huge volume of data, which creates both an opportunity and a need to extract knowledge from databases.

Data mining is a process of nontrivial extraction of implicit, previously unknown and potentially useful information from data [41]. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. Balazs Feil, 2008 [33] stated that the goals of data mining were achieved via clustering, summation, regression and change and deviation detection methods.

Gözde Ulutagay, 2016 [34] stated that uncertainty was a widespread phenomenon in data mining problems. The ongoing challenges of uncertainty give rise to a plethora of knowledge extracting methods that use fuzzy logic. It aimed to present recent outstanding developments and trends in the theory and algorithms of data mining using fuzzy logic.

Jianxiong Luo, 1999 [37] explored integrating fuzzy logic with data mining methods for intrusion detection. The integration with fuzzy logic could produce more abstract and flexible patterns for intrusion detection, since many quantitative features are involved in intrusion detection and security itself is fuzzy. Xie, 2006 [39] proposed fuzzy decision tree and simple fuzzy logic rules to solve both classification and prediction problems for fuzzy logic data mining and machine learning. This approach was used to interpret the information of the tree only if the accuracy of the training set using these rules is reasonably close to the accuracy using fuzzy decision tree.

Corrado Mencar et.al, 2007 [35] described and commented a number of issues that need to be addressed to provide for understandable patterns. A careful consideration of all such issues might end up in a systematic methodology to discover comprehensible knowledge from data. J. Aroba, 2007 [36] applied a set of clustering algorithms based on fuzzy logic and data mining and allowed to obtain data in the form of linguistic rules and charts about the behavior and odiel river estuary affected by acid mine drainage. Z. Qian et al. 2012 [40] devised an algorithm for data mining based on Fuzzy Logic in establishing the eigen set of latent relationship of data. The algorithm was propitious to improve the efficiency of data mining.

Balazs Feil, 2008 [33] aimed to give a comprehensive view about the links between fuzzy logic and data mining. Here, knowledge extracted from simple / huge databases could be represented by fuzzy rule-based expert systems. It was highlighted that both model performance and interpretability of the mined fuzzy models were of major importance. S. Taneja et.al 2016 [38] developed a new algorithm to handle the classification by using fuzzy rules on the real world data set. The proposed algorithm catered in handling admission of students to various Universities by classifying them. They proved that their algorithm was more efficient than others in terms of performance.

SMOTE [3] is the best technique to make the dataset as a balanced compared to various under and oversampling technique. The authors were used to proven a using various classifiers and they showed their result in various recall, precision, f-measure to proven their technique produced the better performance. SMOTE [1] increased the classification performance of various lazy learners, rule-based induction models, and tree-based models. The author showed that the majority of classification techniques performed better over balanced dataset.

SMOTE and ensemble machine learning approach [4] suggested for predicting diabetes mellitus. The authors took diabetes dataset [imbalanced] and applied random under sampling and SMOTE technique into the dataset to make them as balanced. Finally, the resultant dataset is applied to J48, NB tree, Logistics and random forest classifier and proven that SMOTE technique performs better performance over the imbalanced data compared to random under sampling. A study on SMOTE [20] enhanced the accuracy of different induction and decision tree models in order to predict kidney diagnosis of patients is presented. It is concluded that the accuracy of SMOTE on decision tree is better than rule-based models.

In a modified Principal Component Analysis [43] is presented for symmetric fuzzy data. According to Krol, 2006 [44], they deal with models on the basis of fuzzy data. There are algorithms to cluster, to classify or to visualize fuzzy data [45].

A comparative analysis using cardiotocography data is [2] presented. It enhanced the classification performance of classifier using feature selection and normalization technique. They are using J48, IBK, Logistic, SMO, Random Forest and Naive Bayes classifier in their experimentation work. A study [27] on numerous data mining classification techniques is presented. It included the genetic algorithm, KNN, SVM, C4.5, CART, etc., the pros and cons of each algorithm was described. A study on classification algorithms [10] on crime and accident in a city of USA. They analyzed five classification algorithms such as JRIP, Naïve Bayes, J48, BayesNet, and Decision Table. It is concluded that the JRIP and decision table provides the highest accuracy.

It is proposed a new under sampling algorithm [5], in that they used different strategy to select nearest neighbors from the majority class and proven that it can provide better classification performance over various classifier.

III. PROPOSED ALGORITHM

The entire knowledge available in a high dimensional imbalanced dataset is not always necessary to define various categories represented in the dataset. Though the ML and DM techniques are suitable for handling data mining problems, they may not be effective for handling high dimensional imbalanced data. When classifier deals with imbalanced dataset it did not classify properly to predict the output and another major issues in that is high dimensionality over the dataset. When the dimension of the input data increases, the accuracy and efficiency of the result produced by the DM algorithm will decreases rapidly. This inspires the need for the efficient feature selection process in the area of data mining. The process of feature selection transforms the original high dimensional feature vector into low dimensional feature vector. The main objective of feature selection is to reduce the high dimensional of the original dataset and to improve the mining performance such as predictive accuracy, speed of learning, etc.

3.1 Visualization of Data

Since in practical DM problems high dimensional data have to be dealt with, in most of the cases it would be very useful if we could see the structure of these data in a low dimensional space. The reduction of dimensionality of the feature space is also important because of the curse of dimensionality. In a nutshell, same number of examples fills more of the available space when the dimensionality is low, and its consequence is that exponential growth with dimensionality in the number of examples is required to accurately estimate a function.

Two general approaches for dimensionality reduction are feature extraction and feature selection. *Feature extraction* includes transforming the existing features into a lower dimensional space, and *feature selection* includes selecting a subset of the existing features without a transformation. Feature extraction means creating a subset of new features by combination of existing features. These methods can be grouped based on linearity. A linear feature extraction or projection expresses the new features as linear combination of the original variables. The type of linear projection used in practice is influenced by the availability of category information about the patterns in the form of labels on the patterns. If no category information is available, the Eigenvector projection (or PCA) is commonly used.

Principal Component Analysis (PCA) takes a data set $X = [x_1, x_2, \dots x_n]^T$ where $x_k = [x_{1,k}, x_{2,k,} \dots x_{n,k}]^T$ is the k^{th} sample or data point in a given orthonormal basis in and finds a new orthonormal basis $U = [u_1, u_2, \dots u_n]$ where $u_i = [u_{1,i}, x_{2,i,} \dots x_{n,i}]^T$ with its axes ordered. This new basis is rotated in such a way that the first axis is oriented along the direction in which the data has its highest variance. The second axis is oriented along the direction of maximal variance in the data, orthogonal to the first axis. Similarly, subsequent axes are oriented so as to account for as much as possible of the variance in the data, subject to the constraint that they must be orthogonal to preceding axes. Consequently, these axes have associated decreasing 'indeces', λi, i=1,2,…n, corresponding to the variance of the data set when projected on the axes. The principal components are the new basis vectors, ordered by their corresponding variances. The vector with the largest variance corresponds to the first principal component.

And another major issues in that is, imbalanced class distribution problem. While the Classification and prediction is highly affected because of the imbalanced class problem. Classifier may misclassify over the majority class instances, it will ignore the minority class instances. It may leads to the severe drawback while diagnosis the very rarely predicted diseases and it will definitely misprediction while diagnosis. Hence, it is introduced an effective hybrid approach to overcomes these issues. It combines feature selection technique (PCA) with SMOTE. PCA does dimensionality reduction and SMOTE balances the imbalanced class distribution.

In dataset each objects is classified based on its similarities. In DM, the main aim of the classification algorithm is to accurately predict the target class of objects. In our experimentation study, we used Lazy learners (IBk, Kstar), Rule based (Jrip, Ridor), Tree based (J48, NB Tree) algorithms. These algorithms are currently widely used in medical diagnosis. Table 1 describes the entire algorithms used in this study.

Table I: Algorithms used

Category	Algorithms
Lazy learners	IBk, Kstar
Rule based	Jrip, Ridor
Tree based	J48, NB Tree

IBK denotes Instance based k-nearest Neighbors algorithm and it is developed [7]. It is one of the type of lazy learner, which implements the K-nearest neighbor algorithm, as if it generates a model for classification at the testing phase, Kstar is an instance-based classifier and it is a lazy learner. Jrip is also said to be RIPPER (Repeated incremental pruning to produce error reduction). It is one of the most popular ruled based algorithm. It usually divides the dataset into classes and generates rules includes all the attribute of the class and same process with all other classes. Another rule-based algorithm it is used in this research is Ridor. Ridor denotes Ripple Down Rule learner and it works in two phases. Initially the default rules are constructed, in phase two; exceptions are produced for the rules generated by default with lowest error rate.

In tree based, we used J48 and Naive Bayes classifier. J48 is also known as ID3 algorithm and it generates the decision tress at the time of classifying the data. Naïve Bayes classifier is the widely used classifier and it is developed in [7]. It consist of conditional probabilities. It usually needs less training data to build a model.

The proposed framework deals with the high dimensional imbalanced class distribution problem. The main objective of this work is to increase the classification accuracy of the various classifier. The proposed work consist of two main techniques. The functionality of the proposed framework is detailed in Fig.1.

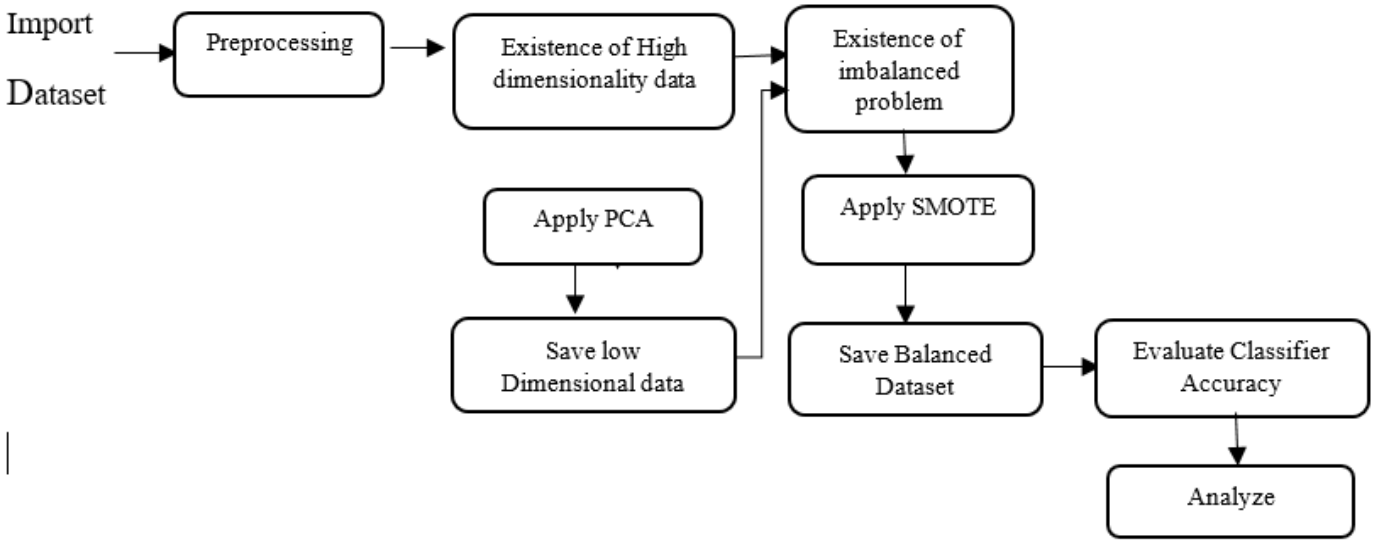

Fig.1. Functionality of the Proposed Model

Initially the raw data is taken from repository, and then preprocessed the raw data using Weka datamining tool [1]. The preprocessed dataset is in .arff format. Then if the dataset have the existence of high dimensionality, we will perform PCA else check whether it have imbalanced class problem or not. If yes, apply SMOTE then evaluate the classifier accuracy and analyze the result using various metrics. The detailed description of SMOTE and PCA will be seen in following sub-section and the algorithm of the proposed model is given as follow

3.2 Methodology

This section focus the suggested approaches for conducting the study. The main aim of this work is to increase the accuracy of the classifier over the medical datasets. This work consist of three main stages of process. The first stage is to calculate the accuracy of original datasets. The second stage is to apply PCA over the dataset and retrieve the accuracy and the third stage is to apply SMOTE over the retrieved dataset from PCA. Finally, the resultant accuracy will shows the overall performance of the various classifier after applying SMOTE along with PCA.

Algorithm 1: Proposed Model

1. Procedure Model()
2. **INPUT** = Medical dataset
3. **OUTPUT** = Synthetic Minority Oversampling technique with PCA

4. If high dimensional data exists
5. **DO**
6. PCA feature Selection Technique
7. **ELSE**
8. Check for balanced data class
9. **IF**(dataset!=balanced data)
10. **THEN**
11. Apply SMOTE
12. Calculate the accuracy of the dataset
13. Evaluate the model
14. **End**

3.2.1 SMOTE

SMOTE alogirthm [3, 17, 23, 26, 29, 32] transforms the imbalanced class distribution to the balanced class distribution [3]. This approach is motivated by a technique which proven to be successful in the recognition of handwritten characters [11].Various oversampling and under sampling technique is available to address the imbalance class problem but that is does not much effective when compared to SMOTE. While dealing with imbalanced class problem, SMOTE synthetics a new artificial minority class instance [29] and match up with the majority class sample rather than doing under sampling or oversampling without replacement. With the following steps, SMOTE creates a new synthetic minority samples:

1. Initially, take the difference between a feature vector (minority class sample) and one of its k nearest neighbors (minority class samples).
2. Then, multiply this difference by a random number between 0 and 1.
3. Finally, add this difference to the feature value of the original feature vector, thus a new feature vector is created [13].

SMOTE creates an artificial instance is showed below with an example. Assume a feature vector (7, 3) and its nearest datapoint is (5, 4). Let:

F_{11} is the first feature value of first data point.

F_{21} is the first feature value of second data point.

F_{12} is the second (nearest) feature value of first data point

F_{22} is the second (nearest) feature value of second data point.

Perform the $F_{new1} = F_{21}-F_{11}$

$F_{new2}=F_{22}-F_{12}$

As a result

$F_{new1} = (5-7) = -2$; $F_{new2} = (4-3) =1$;

The below given formula will be used to generating the new instance:

$(F_{new1}, F_{new2}) = (F_{11}, F_{12}) + \text{Rand } (0\text{-}1)*(F_{new1}, F_{new2})$

$(F_{new1}, F_{new2}) = (7, 3) + \text{Rand } (0\text{-}1)*(-2, 1)$

$= (7, 3) + (0.2)*(-2, 1)$

$= (7, 3) + (-0.4,-0.2)$

$= (6.6, 2.98)$ is the newly created datapoint of the taken example.

Rand (0-1) will randomly generate the number between 0 to 1.
The resultant shows the newly synthetic data point and the below example shows how exactly SMOTE works to create the artificial sample. The same procedure will follows until reaches the minority classes matches the majority classes. Though SMOTE produce the instance artificially but it does not much practically effective on high dimensional data. So we applied PCA along with SMOTE to overcome the high dimensionality imbalanced data problem.

3.2.2 Principle Component Analysis

In data mining and machine learning, classification algorithms always suffers while dealing with high dimensional data. Traditional classification algorithms can achieve better performance in low-dimensional data, and have poor performance with high-dimensional data [15]. For example, when classification algorithm working with text classification and recognitions systems, data usually can hold thousands or even millions of dimensions. It does directly work with the original data; a model typically comes up that is so complex that it is possible to occur overfitting problem [25].

In addition, the inconsistency and noise interference that high-dimensional data cannot even get rid of increases computational complexity and extend the training time. Hence, the computational complexity of high-dimensional data must be reduced which will increase the efficiency of the classification algorithm [12]. So we used PCA along with SMOTE. It is among the most popular feature extraction method. It is mainly for dimensionality reduction in which it transforms data from higher dimension to lower dimension. It is a way of identifying

patterns in data and expressing the data in such a way to highlight their similarities and differences. The following example show how exactly PCA deal with high dimensional data shown in Table II.

Table II: Algorithms used

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2.0	1.6
1	1.1
1.5	1.6
1.1	0.9

Let us consider each X and Y values are attribute value of data point.

Step 1: Perform mean on both X and Y

$\bar{X}=\frac{\sum_{i=0}^{n} fX}{n}$ $\bar{Y}=\frac{\sum_{i=0}^{n} fY}{n}$

Mean value of x=1.81 and Mean value of y=1.91

Step 2: Subtract the original X, Y value to the mean value, so that to make the data pass through the origin. This process is known as Data Adjust. The resultant of data adjust will be shown in ***Table III***.

Table III: Resultant of Data Adjust

X	Y
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.7	-1.01

The graph for the data adjust points, be like the following graph

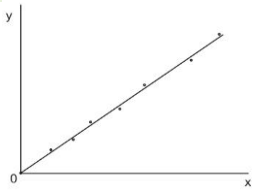

Figure 2. Graph for Data Adjust points

The mean of the data adjust resultant value would be zero, that is why the line will pass through the origin. It is noted that when plot the original X, Y data, the line will not directly pass through the origin, it will deviate from the origin so that we will done data adjust.

Step 3: Calculate the Covariance Matrix: Covariance Matrix is a measure between two dimensions. They show how two variables vary together.

$\text{Cov}(X, Y) = \frac{\sum_{i=0}^{n}(X-\bar{X})(Y-\bar{Y})}{n-1}$

X denotes the original data point of x-axis

$\bar{X}$ denotes the mean of X

Y denotes the original data point of y-axis

$\bar{Y}$ denotes the mean of Y

N denotes number of observations

The following example shows the demonstration of covariance matrix. Let us take,

X	2.1	2.5	3.6	4.0
Y	8	10	12	14

Here, $\bar{X}$=3.1; $\bar{Y}$=11; n=4

$$Cov(X,Y) = \frac{\sum_{i=0}^{n}(2.1-3.1)(8-11)+(2.5-3.1)(10-11)+(3.6-3.1)(12-11)+(4-3.1)(14-11)}{3}$$

= 2.26

Note: we got positive value, which means that it would vary in the same direction(X increases, y also increases).

Step 4: What about Covariance Matrix?

	X	Y
X	cov(X, X)	cov(X, Y)
Y	cov(Y, X)	cov(Y, Y)

The covariance matrix for the data considered is given as

Cov(X, Y) =

	X	Y
X	0.616	0.6154
Y	0.615	0.7165

Since the non-diagonal elements in this covariance are positive, it can expect that both X and Y variable increases together.

Step 5: Calculate the Eigen vectors and Eigen values for the covariance Matrix.

Eigen vector is a projected vector from the data. It is normally perpendicular to the data or new direction in which the most significant data lies.

Eigen values = $\begin{pmatrix} 0.4908 \\ 1.25402 \end{pmatrix}$ Eigen vectors = $\begin{pmatrix} -0.735 & -0.678 \\ 0.677 & -0.731 \end{pmatrix}$

The Eigen value 0.4908 will corresponding to the first column values of the Eigen vectors and the Eigen value 1.25402 will corresponding to the second column values of the Eigen vector. The most important (principal) Eigen vector would have the direction in which the variables strongly correlate.

Step 6: The Eigen vectors with highest Eigen value will be chosen as the PCA, and then we ignore rest of the dimension.

i. Given the n dimensions of the data(here n=2), so we will get 'n' Eigen vector
ii. Then choose P Eigen vector, where p<n, hence we reduce the original dimension

So now, we re-represent the entire data using the following formula (Note: The data we taken here is data adjust points)

Final data= Row feature vector * Row data adjust

Row feature vector.is the matrix with the Eigen vectors in the column transposed, so that they are now in rows. *Row data adjust* is the mean adjusted data transposed (i.e.): the data items are in each column with each row holding a separate dimension. *Final data* is the final dataset, with data items in columns and dimensions along with rows

The original data had two axes [x, y]. Therefore, our data was in terms of them. Now they are in terms of Eigen vectors. The new dataset would have reduced its dimension, if we have chosen to cut an Eigen vector. The other transformation we can make is by taking only the Eigen vector with the largest Eigen value (Dimensionality reduction again).

IV. EXPERIMENT AND RESULTS

The proposed system is implemented using Windows 10 Operating Systems, Intel® Core™ i5 processor. For training and testing, the dataset is divided into 2:3 and 1:3 ratios respectively Core i5 processor having 12GB RAM capacity and in Java language. For evaluating the algorithms under consideration, it is used Cardiotocograms data from UCI Machine Learning Repository. The algorithms used in this dataset were extracted from 'WEKA' packages and incorporated into NetBeans. The suggested SMOTE along with PCA technique is designed in java and implemented over the algorithms.

For this study, it is gathered dataset from UCI machine learning repository [ ]. It is used two types of datasets for this experiment. The first dataset belongs to the Cardiotocography and it has the measurements of FHR and uterine contraction (UC) features on CTG classified by expert obstetricians. Table IV gives the description of dataset. In this section the dataset description and the applied methodology is discussed. Initially the dataset has 2126 instances with 23 attributes. The first has 295 instances, the second class has 1655 instances and third class has 176 instances.

The CTGs were classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C. ...) to a fetal state (N, S, and P). Therefore the dataset can be used either for 10-class or 3-class experiments. In this project, 3-class experiment is used. From the number of instances it is studied that the classes are not balanced. To overcome this difficulties the dataset must be made to balanced, that is synthetic instances are to be created to get a balanced dataset.

Table IV: Carditocography Dataset description

S.NO	ATTRIBUTE NAME	TYPE
1	LB	Real
2	AC	Real
3	FM	Real
4	UC	Real
5	DL	Real
6	DS	Real
7	DP	Real
8	ASTV	Real
9	MSTV	Real
10	ALTV	Real
11	MLTV	Real
12	Width	Real
13	Min	Real
14	Max	Real
15	Nmax	Real
16	Nzero	Real
17	Mode	Real
18	Mean	Real
19	Median	Real
20	Variance	Real
21	Tendency	Real
22	FHR pattern class code	Real
23	Fetal state class code	categorical

Summary of Carditocography Dataset is as follows.

Instances: 2126, #Attributes: 23, #Classes: 3 (One, Two, Three).

Attribute Characteristics: Real, Associated Tasks: Classification.

The performance of machine learning algorithms is typically evaluated by a confusion matrix as illustrated in Table V (for a 2-class problem). The columns are the Predicted class and the rows are the Actual class. In the confusion matrix, TN is the number of negative examples correctly classified (True Negatives),

Table V: Confusion Matrix

	Predicted Negative	Predicted positive
Actual Negative	TN	FP
Actual Positive	FN	TP

FP - the number of negative examples incorrectly classified as positive (False Positives), FN is the number of positive examples incorrectly classified as negative (False Negatives) and TP is the number of positive examples correctly classified (True Positives). Predictive accuracy is the performance measure generally associated with machine learning algorithms. The evaluation metrics used in this work are,

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (1)$$

$$Precision = \frac{tp}{tp+fp} \quad (2)$$

$$Recall = \frac{tn}{tn+fp} \quad (3)$$

$$F-measure = \frac{2*Precision*recall}{tn+fp} \quad (4)$$

For Evaluation of imbalanced and balanced dataset, the algorithms like Jrip, Ridor, J48, Naive Bayesian Tree, IBK and K- star are used. The accuracy of the Carditocography data is calculated and it is summarized in the following tables below. The algorithm listed in **Table VI** is applied with both SMOTE and PCA techniques are applied and their respective results were listed in below following tables.

In the first stage, we applied SMOTE to the dataset, then we compare the accuracy of original dataset and SMOTE applied dataset of various classifiers used and it is shown in table 6 and the comparison graph is given in figure 3. It is found that IBK produce better accuracy in both original dataset and SMOTE applied dataset.

Table VI: Accuracy of classifier between original dataset and SMOTE applied dataset over Cardiocography

	Original Dataset				SMOTE			
Classifier	**Accuracy**	**Precision**	**Recall**	**f-measure**	**Accuracy**	**Precision**	**Recall**	**f-measure**
IBK	**98.96**	0.99	0.99	0.99	**99.21**[&]	0.992	0.992	0.992
Kstar	94.21	0.941	0.942	0.941	94.52	0.945	0.945	0.945
Jrip	98.63	0.986	0.986	0.986	98.87	0.989	0.989	0.989
Ridor	97.69	0.977	0.977	0.977	98.52	0.985	0.985	0.985
J48	98.68	0.987	0.987	0.987	98.61	0.986	0.986	0.986
NB Tree	90.59	0.921	0.906	0.911	94.65	0.951	0.947	0.948

[&] - Highest accuracy indicates in terms of bold.

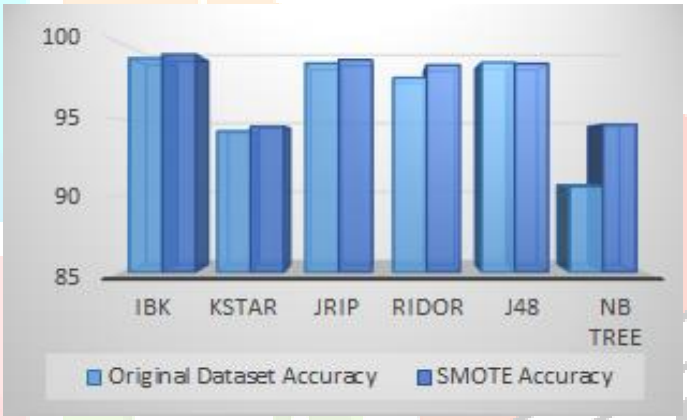

Fig.3: Comparison of accuracy with original dataset and SMOTE applied dataset

In the second stage, we applied SMOTE along with PCA over the dataset. Since PCA reduce the dimensionality of the data set, it can able produce better accuracy. The results are compared with SMOTE and it is listed in Table VII and comparison graph is given in figure 4. It shows that SMOTE with PCA produce better accuracy than SMOTE except for Jrip and Ridor algorithm.

Table VII: Accuracy of classifier between SMOTE and SMOTE along with PCA over cardiocography

	SMOTE				SMOTE with PCA			
Classifier	**Accuracy**	**Precision**	**Recall**	**f-measure**	**Accuracy**	**Precision**	**Recall**	**f-measure**
IBK	**99.21**	0.992	0.992	0.992	99.55	0.996	0.996	0.996
Kstar	94.52	0.945	0.945	0.945	**99.67**[&]	0.997	0.997	0.997
Jrip	98.87	0.989	0.989	0.989	98.51	0.985	0.985	0.985
Ridor	98.52	0.985	0.985	0.985	98.16	0.982	0.982	0.982
J48	98.61	0.986	0.986	0.986	98.65	0.986	0.986	0.986
NB Tree	94.65	0.951	0.947	0.948	96.98	0.971	0.970	0.970

[&] - *Highest accuracy indicates in terms of bold.*

Fig.4: Comparison of accuracy with SMOTE and SMOTE with PCA applied dataset

The final stage is to compare the accuracy of algorithms over original dataset and after applied SMOTE with PCA dataset. The results are listed in Table VIII and also the corresponding graph is given in figure 5. It is seen that balanced and low dimensional dataset (i.e.) of the proposed SMOTE with PCA produce better accuracy than the others

Table VIII: Results

	Original Dataset				**SMOTE with PCA**			
Classifier	**Accuracy**	**Precision**	**Recall**	**f-measure**	**Accuracy**	**Precision**	**Recall**	**f-measure**
IBK	**98.96**	0.99	0.99	0.99	99.55	0.996	0.996	0.996
Kstar	94.21	0.941	0.942	0.941	**99.67**[&]	0.997	0.997	0.997
Jrip	98.63	0.986	0.986	0.986	98.51	0.985	0.985	0.985
Ridor	97.69	0.977	0.977	0.977	98.16	0.982	0.982	0.982
J48	98.68	0.987	0.987	0.987	98.65	0.986	0.986	0.986
NB Tree	90.59	0.921	0.906	0.911	96.98	0.971	0.970	0.970

[&]-Highest accuracy indicates in terms of bold.

Figure 5: Comparison of accuracy with Original Dataset and SMOTE with PCA applied dataset

In Fig.3,4,5 shows the overall comparison of the various technique used for the experimentation. In that X-axis represent the various classifier used, Y-axis represents the percentage of accuracy. After analyzing the various classifier performance over SMOTE and our proposed SMOTE with PCA, it is found that our approach (SMOTE with PCA) performs better over the imbalanced dataset, produce the better accuracy, and visually represents that in Fig.3. Lazy learner Kstar classifier is recorded better performance with our approach and it has **99.67%** accuracy.

V. CONCLUSION & FUTURE ENHANCEMENTS

Data mining is a process of discovering useful patterns from a large set of data. Data mining is mostly used in large information processing applications including Healthcare. Classification technique of data mining classifies the data into a set of classes based on some attributes for further processing. A hybrid algorithm to handle the classification by using fuzzy data mining on the Healthcare data set.

The Cardiotocography dataset for classification of fetal state class was analyzed using Jrip, Ridor, J48, NBStar, IBk, and Kstar classifiers. Due to huge amount of data, the data are imbalanced and to balance the data SMOTE and proposed SMOTE methodology along with PCA are used, in which it is identified that SMOTE with PCA provides more balanced data set than SMOTE. The results showed that the classification performance on balanced and low dimensionality data provides enhanced performance when compared to imbalanced dataset and SMOTE produce dataset. In future, this technique can be applied in case of Big data with Map Reduce technique, as the traditional mechanism does not fit to handle Big Data.

REFERENCES

[1]. Sai Prasad Potharaju, M Sreedevi, et al. Data mining approach for accelerating the classification accuracy of Cardiotocography. Clinical Epidemiology and Global Health, 7(2):160–164, 2019.

[2]. Razman Afridi, Zafar Iqbal, Muzammil Khan, et al. Fetal Heart Rate classification and comparative analysis using cardiotocography data and known classifiers, Intel. Journal of Grid and Distributed Computing, 12(1):31–42, 2019.

[3]. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, W Philip Kegelmeyer. SMOTE: Synthetic Minority over-sampling Technique, Journal of artificial intelligence research, 16:321–357, 2002.

[4]. Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian, Clinton Brawner, Jonathan Ehrman, Sherif Sakr. Predicting Diabetes Mellitus using smote and Ensemble Machine Learning Approach: The henry ford exercise testing (fit) project PloS one, 12(7), 2017.

[5]. Małgorzata Bach, Aleksandra Werner, and Mateusz Palt. The proposal of under sampling method for Learning from Imbalanced Datasets. Procedia Computer Science, 159:125–134, 2019.

[6]. Chumphol Bunkhumpornpat, Krung Sinapiromsaran, Chidchanok Lursinsap, Safe-level-SMOTE: Safe-level-synthetic Minority Over-sampling Technique for Handling the Class Imbalanced Problem. Pacific-Asia Conf. on knowledge discovery and data mining, 475–482. Springer, 2009.

[7]. Safae Sossi Alaoui, Yousef Farhaoui, Brahim Aksasse. Classification Algorithms in Data Mining. 2018.

[8]. Liang-Hwa Chen, Shyang Chang. An adaptive learning algorithm for principal component analysis. IEEE Transactions on Neural Networks, 6(5):1255–1263, 1995.

[9]. Deepthy K Denatious, Anita John. Survey on data mining techniques to enhance intrusion detection. In 2012 Intel. Conf. on Computer Communication and Informatics, 1–5, 2012.

[10]. Amit Gupta, Azeem Mohammad, Ali Syed, Malka N Halgamuge. A comparative study of classification algorithms using data mining: crime and accidents in Denver city the USA Education, 7(7):374–381, 2016.

[11]. Thien M Ha, Horst Bunke. Off-line, Handwritten Numeral Recognition by perturbation method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5):535–539, 1997.

[12]. Qi Hang, Jinghui Yang, Lining Xing. Diagnosis of rolling bearing based on classification for high dimensional unbalanced data. IEEE Access, 7:79159–79172, 2019.

[13]. Shengguo Hu, Yanfeng Liang, Lintao Ma, Ying He. Msmote: Improving classification performance when training data is imbalanced. IEEE international workshop on computer science and engineering, 2, 13–17. 2009.

[14]. A Kautkar Rohit. A comprehensive survey on data mining.

[15]. Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, Zhigang Gao. A Hybrid Feature Selection Algorithm for Gene Expression Data Classification. Neurocomputing, 256:56–62, 2017.

[16]. Lara Lusa et al. SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics, 14(1):106, 2013.

[17]. Nadir Mustafa, Jian-Ping Li. Medical Data Classification Scheme based on Hybridized Smote Technique (HST) and Rough Set Technique (RST). IEEE Intel., Conf., on cloud computing and big data analysis, 49–55. 2017.

[18]. S Neelamegam, E Ramaraj. Classification algorithm in data mining: An overview. International Journal of P2P Network Trends and Technology (IJPTT), 4(8):369–374, 2013.

[19]. Sai Prasad Potharaju, M Sreedevi. Ensembled Rule based Classification Algorithms for Predicting Imbalanced Kidney Disease Data, Journal of engineering science and technology review, 9(5):201–207, 2016.

[20]. Sai Prasad Potharaju and M Sreedevi. An Improved Prediction of Kidney Disease using smote: Indian Journal of Science and Technology, 9(31):1–7, 2016.

[21]. Delveen Luqman Abd Al-Nabi, Shukri Shereen Ahmed. Survey on Classification Algorithms for Datamining (comparison and evaluation), Computer Engineering and Intelligent Systems, 4(8):18–24, 2013.

[22]. Yoga Pristyanto, Irfan Pratama, Anggit Ferdita Nugraha. Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multi Class Classification. IEEE Intel. Conf. on Information and Communications Technology (ICOIACT), 310–314, 2018.

[23]. Rishabh Rustogi, Ayahs Prasad. Swift Imbalance Data Classification using Smote and Extreme Learning Machine. International Conference on Computational Intelligence in Data Science (ICCIDS), IEEE, 1–6, 2019.

[24]. Foued Saˆadaoui, Pierre R Bertrand, Gil Boudet, Karine Rouffiac, Fr´ed´eric Dutheil, Alain Chamoux. A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining. IEEE transactions on Nano bioscience, 14(7):707–715, 2015.

[25]. Durmus¸Ozkan S¸ Ahin, Nurullah Ates¸ Erdal Kilic, Feature selection in text classification. IEEE 24th Signal Processing and Communication Application Conference (SIU), 1777–1780. 2016.

[26]. Phakhawat Sarakit, Thanaruk Theeramunkong, Choochart Haruechaiyasak. Improving Emotion Classification in Imbalanced YouTube Dataset using Smote Algorithm. Intel. Conf. on Advanced Informatics: Concepts, Theory and Applications, IEEE, 1–5. 2015.

[27]. M Sujatha, S Prabhakar, G Lavanya Devi. A Survey of Classification Techniques in Data Mining. International Journal of Innovations in Engineering and Technology (IJIET), 2(4):2319–1058, 2013.

[28]. S Umadevi, KS Jeen Marseline. A Survey on Data Mining Classification Algorithms. Intel. Conference on Signal Processing and Communication (ICSPC), IEEE, 264–268, 2017.

[29]. Yuanting Yan, Ruiqing Liu, Zihan Ding, Xiuquan Du, Jie Chen, Yanping Zhang. A parameter-free cleaning method for smote in imbalanced classification. IEEE Access, 7:23537–23548, 2019.
[30]. Wen-Hui Yang, Dao-Qing Dai, Hong Yan. Feature Extraction and Uncorrelated Discriminant Analysis for High-dimensional Data. IEEE transactions on knowledge and data engineering, 20(5):601–614, 2008.
[31]. Zhengwu Yuan, Pu Zhao. An Improved Ensemble Learning for Imbalanced Data Classification. In2019, IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, IEEE, 408–411, 2019.
[32]. Min Zeng, Beiji Zou, Faran Wei, Xiao Liu Lei Wang. Effective Prediction of Three common Diseases by Combining Smote with Tomek Links Technique for Imbalanced Medical Data. IEEE Intel. Conf. of Online Analysis and Computing Science (ICOACS), 225–228. 2016.
[33]. Balazs Feil, 2008. Janos Abonyi , Introduction to Fuzzy Data Mining Methods, Ph.D. Thesis, University of Pannonia, Dept. of Process Engg., Hungary.
[34]. Gözde Ulutagay, Ronald Yager, Bernard De Baets, Tofigh Allahviranloo, 2016. Forefront of Fuzzy Logic in Data Mining: Theory, Algorithms, and Applications, Hindawi Publishing Corporation, Advances in Fuzzy Systems,1-2.
[35]. Corrado Mencar, Giovanna Castellano, Anna M. Fanelli, (2007). On The Role of Interpretability in Fuzzy Data Mining, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15(5), 521-537.
[36]. J. Aroba, J. A. Grande J. M. Andu´ Jar, M. L. de la Torre J. C. Riquelme, Application of fuzzy logic and data mining techniques as tools for qualitative interpretation of acid mine drainage processes, Springer-Verlag, 2007.
[37]. Jianxiong Luo, Integrating Fuzzy Logic With Data Mining Methods for Intrusion Detection, 1999.
[38]. S. Taneja, B. Suri, H. Narwal, A. Jain, A. Kathuria and S. Gupta, "A new approach for data classification using Fuzzy logic," 6th International Conference - Cloud System and Big Data Engineering (Confluence), pp. 22-27, 2016.
[39]. Xie, Dong, Fuzzy logic data mining and machine learning and its applications, Ph.D. Thesis, University of Bristol, Faculty of Engineering, 2006.
[40]. Z. Qian et al. (Eds.): Algorithm for Data Mining Based on Fuzzy Logic, Recent Advances in CSIE 2011, LNEE 124, Springer-Verlag Berlin Heidelberg, 353–357, 2012.
[41]. Hang Chen, Applications of Fuzzy Logic in Data Mining Process, Advanced Fuzzy Logic Technologies in Industrial Applications.
[42]. Rudolf Kruse, Detlef Nauck, Christian Borgelt, Data Mining with Fuzzy Methods: Status and Perspectives.
[43]. Giordani P, Kiers, H.A.L., Principal component analysis of symmetric fuzzy data. *Computational Statistics and Data Analysis*, 45:519-548, 2014.
[44]. Krol, D., Kukla, G.S., Lasota, T. Trawinski, B. *Fuzzy model for the assessment of operators' work in a cadastre information system*. Intel., Conf., on Knowledge-Based and Intelligent Information and Engg., Systems, 2006.
[45]. Bandemer, H.. Mathematics of Uncertainty. Ideas, Methods, Application Problems, *Studies in Fuzziness and Soft Computing*. Springer, Vol.189, 2006.
[46]. Zadeh LA, 1965. Fuzzy sets, information and control, Vol (8), pp.338–353.
[47]. Holsheimer M, Siebes A, 1994. Data mining: the search for knowledge in Databases. Report CS-R9406, CWI Amsterdam.