A CLASS OVERLAPPING DISTANCE BASED APPROACH FOR FEATURE SELECTION FOR CONTINUOUS DATA

Apoorva Yadav*, Ramratan Ahirwal and Shailendra Shrivastava

Department of Computer Science and Engineering, Samrat Ashok Technological Institute, Vidisha, 464001, India

*Corresponding author email: apmanit2018@gmail.com

ABSTRACT

One of the most popular pre-processing techniques in data mining and classification is feature selection. If the dataset is continuous, it might be challenging to pick the most important features from the vast amount of data that is available. This study presents a novel feature selection method for continuous data based on an overlapping class strategy (having normal and non-normal distributions). The classification uncertainty of an instance increases along with the area of class overlap between two classes, and it decreases along with the area. Three possible cases of class overlapping, i.e., complete class overlap, partial class overlap, and no class overlap, are caused by the uncertainty problem resulting from overlapping areas in classification. Therefore, the k-best feature selection in continuous features could make use of this variant overlapping area-based concept. The proposed framework for k-best feature selection employs a normalised overlapping distance and follows the forward feature selection strategy. The approach has been applied to the datasets of ailments such as breast cancer, diabetes, and heart disease. The method produces better outcomes on these datasets, with accuracy rates of 81.36% for heart disease, 76.74% for diabetes, and 95.54% for breast cancer. The experimental data demonstrate that the suggested strategy may be utilised to choose the best features for continuous data, enhance the categorisation of continuous data, and the results provide better accuracy for all the datasets under consideration.

Keywords: Feature Selection, Gaussian Function, K-Fold, Continuous Data, Overlapping Class

INTRODUCTION

The major objectives of feature selection methods are enhancing classification performance and selecting the most relevant properties for data classes. Improved feature selection outcomes may

reduce learning time, increase learning efficiency, and make learning simpler [1]. The adoption of the desired attributes, which will lead to an accurate prediction by the predictive algorithms, is the main challenge with the vast amount of data. The processing of high-dimensional data and the improvement of learning efficiency have both been successfully demonstrated in theory and practice using feature selection. As a result, we can select features based on how well they are categorised [2] and as the computer learning models typically evaluate the efficiency of the feature selection strategy, feature extraction approaches can be broadly divided into two broad categories: supervised feature selection and unsupervised feature selection techniques. Here the main emphasis has been given to supervised techniques such as filter, wrapper, and embedded model. The filter model just takes into account the relationship between the feature and the class name [1]. Numerous filter model approaches to select the best feature include information gain, the chi-square test, Fisher's score correlation Coefficient, Variance Threshold, Mean Absolute Difference, Dispersion Ratio, Mutual Dependence, and Relief. In the wrapper methodology, a subset of features is considered, evaluated, and compared with other combinations. It iteratively employs the subset of features to train the algorithm. The best subset of features is chosen based on the output, and the model is trained again. The wrapper model uses a variety of techniques, including forward selection, backward elimination, bi-direction elimination, exhaustive selection, recursive elimination, etc. Regularisation and tree-based methods are the two embedded-based feature selection techniques. Using these techniques, numerous scholars have been exposed to various concepts for feature selection. Here is a succinct assessment of the literature.

Cai et al. [1] presented a fresh look at feature selection in machine learning, in which authors explain feature selection techniques under supervised and unsupervised learning. Lim and Kim [3] suggested the Unsupervised Feature Selection Method Based on Pairwise Dependence. Their approach is founded on the idea of information gain. Verma et al. [4] proposed an approach by using ant colony optimisation and relative fuzzy entropy; feature selection for real-time intrusion detection systems is performed. A global ideal feature subset for matching the original dataset is produced by this heuristic method for selecting the best feature subset. Their results show that adopting the optimal feature subset significantly reduces processing time and computational expenses by 37.19% while also steadily increasing the detection accuracy of the models by 0.27%. Bashir et al. [5] performed an experiment using various machine learning algorithms and a quick miner tool to execute feature selection methods

on diverse heart disease datasets to improve heart disease prediction. Their results were, Decision tree achieved 82.22% accuracy, Logistic Regression accuracy was 82.56%, Random forest showed an improvement from then previous one which, is 84.17%, and Naïve Bayes accuracy was increased from previous research. It achieved accuracy 84.24%, and the greatest accuracy of SVM after applying Logistic Regression is 84.85%.

Jain et al. [6] presented an analysis of the most up-to-date illness prediction categorisation algorithms and feature selection techniques. The study's main goal was to improve and accelerate the diagnosis of chronic diseases through parallel and adaptive classification systems. New hybrid classification techniques should be developed to increase classifier precision and maximise computational effectiveness of outcomes.Manzoor and Kumar [7] proposed an ANN Classifier-Based Feature Reduced Intrusion Detection System. This study suggests a smart system that first ranks features based on information gain and association. An ANN-based classification system was created, trained on a small dataset, and tested on five distinct subsets of the KDD99 dataset. Results reveal that the technique outperforms rival strategies for both attack and non-attack classes. The approach has improved both detection and false alarm rates overall.

Verma et al. [8] compared feature selection for ensemble data mining to forecast skin diseases. In order to anticipate skin diseases, this study examines various data mining techniques. Classifiers using bagging, AdaBoost, and gradient boosting are used as three ensemble methodologies to increase the precision of machine learning algorithms. Six machine learning classification algorithms are used to categorise the prediction of skin diseases: PAC, LDA, RNC, BNB, NB, and ETC. A feature selection technique is used in conjunction with the Gradient Boosting ensemble strategy on RNC to achieve an accuracy of 99.68% on the dataset for skin diseases. Mishra and Singh [9] focused on dimensionality reduction and proposed the use of feature-space clustering in the FS-MLC (Feature Selection for Multi-Label categorisation) approach. It used Decision Tree as the primary classifier across all techniques. FS-MLC is integrated with the cutting-edge techniques of BR and CC. In this case, BR stands for Binary Relevance and CC for Classifier Chain. Mirzaei et al. [10] delivered a two-stage method for selecting features to extract pertinent elements from the speech of Alzheimer's disease patients in the early stages. They employed the kNN, SVM, and DT with wrapper methods as their three classification techniques. Magesh and Swarnalatha [11] offered DT learning based on clusters (CDTL) method for classifying heart illnesses that makes use of huge partitions based on entropy

and distribution samples. The prognosis of heart disease is established through significant and all aspects in RF performance. The RF classifier obtains an increase in prediction accuracy of 89.30% from 76.70% using our CDTL technique (without CDTL).

Hasan and Yukun [12] explained their work in order to determine the most effective approach for predicting cardiovascular illness. This work proposed a comparative evaluation of the filter-wrapper-embedded feature selection method based on the data set's re-sampling and ANN methodology. The suggested method selects features based on their worth using six well-known ML techniques. Sandhiya and Palani [13] suggested an effective approach for predicting diseases, including breast cancer, diabetes, and heart problems. In this illness prediction system, the Intelligent Conditional Random Field (CRF) on feature selection process and the Linear Correlation Coefficient based Feature Selection (ICRF-LCFS) method algorithm are coupled by utilising the Incremental Feature Selection Algorithm (IFSA). Additionally, a Convolutional Neural Network (CNN) with temporal properties already exists (T-CNN). It aids in raising output levels so that forecasting accuracy is greater than 93%. Nagarajan et al. [14] presented a new modeling of heart disease risk via feature selection and categorisation.

This work aims to create a hybrid GCSA, or genetic crow search algorithm, for feature selection and classification employing deep convolution neural networks. The results demonstrate that the suggested model GCSA performs better than the other feature selection approaches, reahing more than 94% classification accuracy.

Gaussian Normal Distribution:

When working with continuous data, it is typically assumed that the continuous values corresponding to each class are distributed according to the Gaussian distribution. In statistics, the normal distributions are frequently described using Gaussian functions. The graph of a continuous probability distribution is a bell-shaped curve, as shown in Figure 1. The probability density is represented by the area under the curve; given that the distribution is continuous, the function determines the probability that a specific event will occur between any two real number limits. The total area of the normal curve is always equal to one. This calculation is done as the curve moves closer to zero on either side of the average. Each attribute's standard deviation and mean are calculated for the database [15]. The mean and standard deviation are calculated for each class formed by the division of the training data. In order to calculate the likelihood of

continuous data gathering, apply the equation below.

$$P(X=x \mid C=c) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where c= class, x= variable to be classified, μ = mean, σ= standard deviation [16].

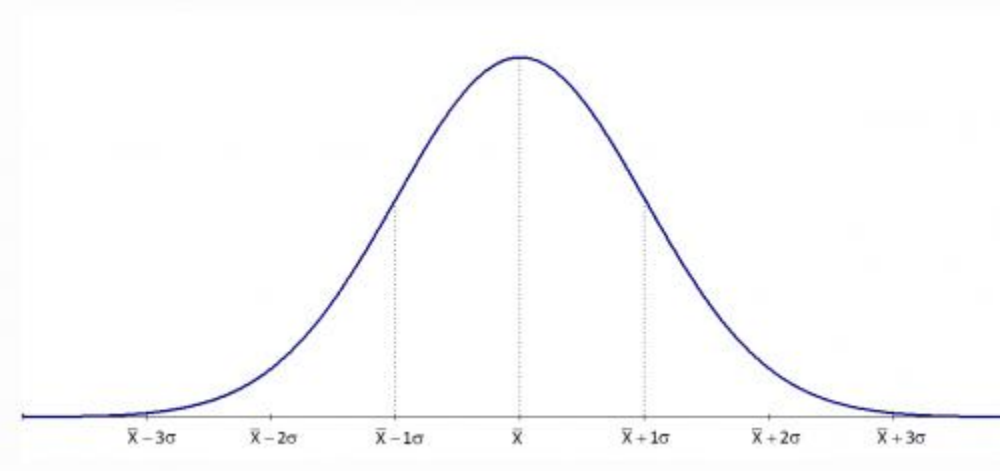


Fig 1: Gaussian Normal Distribution Graph (Bell Curve)

When this Gaussian normal distribution function is used to develop a classifier for continuous data, the concept of an overlapping area between two classes also comes into existence. This concept might be utilised in the selection of k-best features. This varying overlapping area could be of different sizes for different features in the dataset, and each feature class can have three different cases: one of them is complete overlapping between two classes, the second is partial overlapping, and the last could be non-overlapping between classes as shown in Figures 2, 3 and 4.

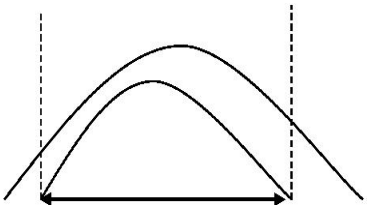


Fig. (2) Class Complete Overlap

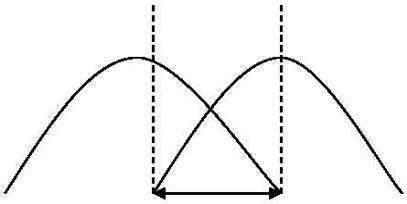


Fig. (3) Class Partial Overlap

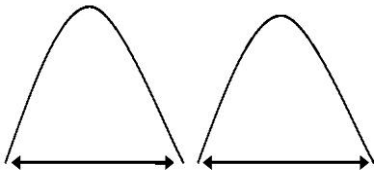


Fig. (4) No Overlap

If the classes are completely overlapped, then the classification of instances becomes more difficult. That means it is difficult to identify the class of a particular instance. Therefore, the feature of this property might be less important in contributing to classification. Therefore, it

could be discarded from the feature selection bucket. If the classes are partially overlapped, then there will be a certain point from which we can accurately classify both classes. Classes that do not overlap indicate there could be a clear-cut boundary, for instance, classification; features with this property could contribute more to the classification of an instance, which plays an important role in the best feature selection approach. The overlapping region or area between two classes is of utmost importance and complicated, so the proposed method replaced it with distance estimation. To pick the k-best attribute, first, the distance will be normalised and brought in between 0 and 1; then, the normalised distance will be arranged in either an ascending or descending order.

For this, a proposed method employed the concept of the intersection of two circles, in which it has been shown that how two circles are touching each other and whether they are touching internally or externally. To show this, the radius and centre's of each circle are required. Two circles will touch if the distance between their centers, d, is equal to the sum of their radii or the difference between their radii, as shown in Figure 5 (a). When two circles satisfy the condition $r1-r2 < d < r1+r2$, as shown in Figure 5 (b), they will intersect. Two circles are concentric when $d = 0$, as shown in Figure 5 (c) [17].

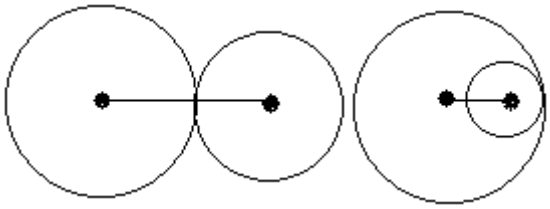


Fig 5 (a): Circles touching externally or internally

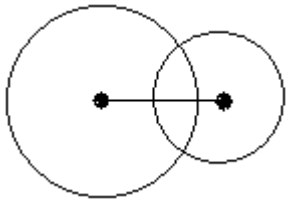


Fig 5 (b): Intersection of two circles

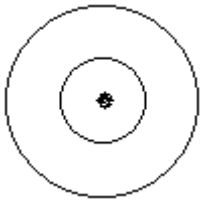


Fig 5 (c): Concentric circles

In light of the aforementioned justifications, the study's key contributions are as follows:

1. The suggested technique classifies continuous data quite effectively.
2. When data is continuous, the features selection method based on class overlapping distance provides the k-best features. It is a completely distinct method for choosing the k-best features.
3. It is simple to use, can be used to big datasets, and produces superior results when compared to other techniques.

The rest of this part discusses the proposed system. It explains the notion or idea

underlying our method and how the best k-features are chosen using the suggested feature selection process, which is addressed in Section 2. Section 3 is the experimental results of this work, which includes discussions of analysis, dataset descriptions, and the final outcomes of our technique. The conclusion of this work and a discussion of prospective applications for this research have been discussed in Section 4.

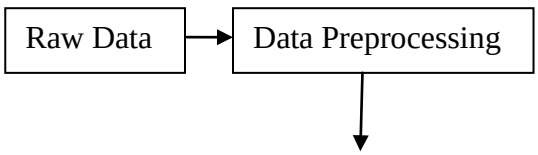
PROPOSED SYSTEM

One of the essential preparation principles in machine learning is feature selection, which significantly impacts the model's performance. So, developing an efficient method for selecting them and choosing the appropriate features becomes challenging every time. To fulfill the need, we developed a unique approach based on the concept of class-overlapping areas for continuous features. The distribution of continuous features could be normal or non-normal, so the approach should also be applicable to both distributions. For normal continuous data distribution, two parameters are evaluated: the mean (μ) and standard deviation (σ). Where mean (μ) is the location pointer that determines the distribution's centre while standard deviation (σ) is the scale parameter that specifies the range of the distribution, it also expresses the dispersion of each observed value from the mean. In our approach, we determine the median and range of both classes to get the location point and the overall range using the standard formula given below-

$$Mean\ (\mu) = \frac{\sum x}{n} \text{ ---------- (1)}$$

$$Standard\ deviation\ (\sigma) = \sqrt{\frac{\sum (x-\mu)^2}{n-1}} \text{ ---------- (2)}$$

Since the area is being discussed in our approach for deciding features and if the distribution of data is normal, this area can be calculated using the Gaussian distribution function. But if the data distribution is not normally distributed, then it is not possible to calculate an accurate area using GNF. As a result, we developed a distance-based approach that is nearly identical to the normal distribution approach and applicable to both types of distributions. The overlapping distance parameter is considered to be the same as the area parameter. Figure 6 delineates a flow chart of the suggested strategy.



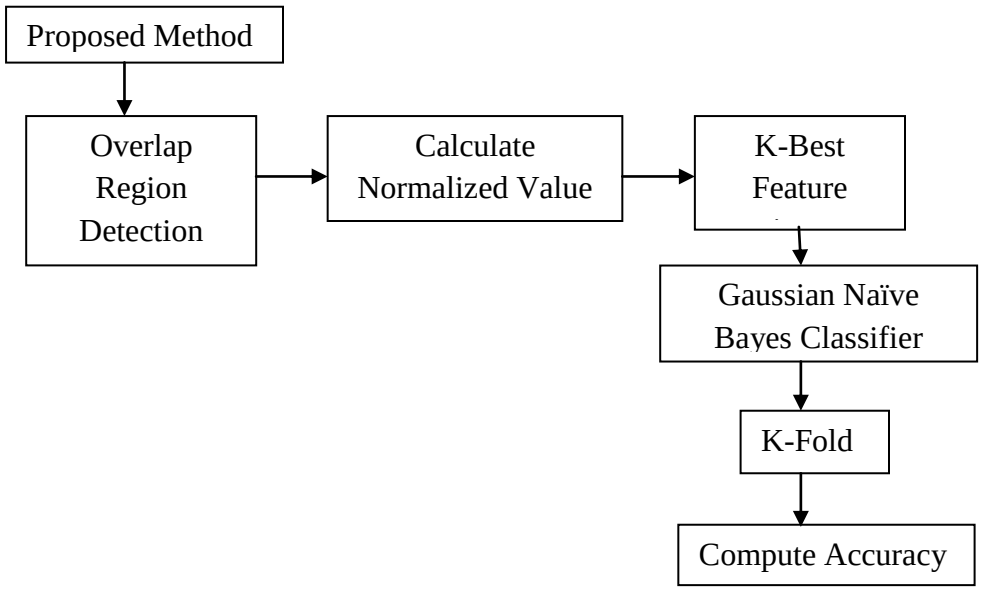


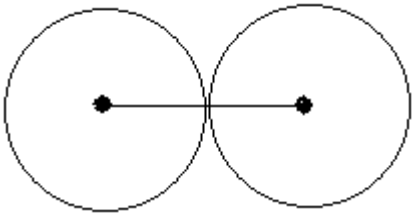
Fig 6: Flow Chart of Proposed System

Overlapping Distance Estimation and Feature Selection Algorithm:

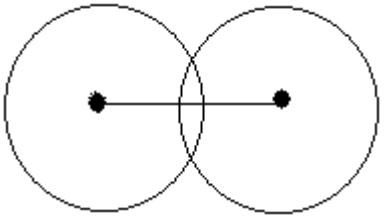
Whenever a continuous variable follows the normal distribution, the normal variate $= \frac{|x-\mu|}{\sigma}$, is used to determine whether a random value is significant or not. If the normal variate variable value finds out that z = 3, that means x ϵ S $(sample)$ with 99.73% confidence, and if z = 4, that means with 99.9% confidence. If the parameters μ, σ are known, then x will fall within the range of μ-zσ <= x <= μ+zσ with varying confidence levels depending on z (if z=3, 99.73% confidence that x ε S). In such a case, class minimum value will be equal to the range min value (μ-zσ), and the maximum value will be equal to the range max value (μ+zσ). So, the two classes have separate minimum and maximum values. The overlapping distance between two classes gives three possible cases, as mentioned above in Figures 2, 3, and 4. In one case, the overlapping distance would be zero or negative when two classes are non-overlapped and calculated using the min and max values of classes, but when the overlapping distance is greater than zero, that means classes are overlapping, so they might be partially overlapped or completely overlapped.

Therefore, the problem is to find out this overlapping distance, which would vary for all features. The problem has been solved by assuming that classes represent μ1 and μ2 as circle

centers. The algorithm we designed is based on the concept of a circle intersection. It may be possible that two circles are touching each other, whether they are touching internally or externally. So to identify this, we need to work on the radius and centre of each circle. The circles will externally touch if the total radius and the distance between the centres are identical, as shown in case 1. Internal interaction takes place when the radius difference and distance between the centres are equal. When the distance between the centres of two circles is greater than their difference in radii but less than their sum, as shown in Case 2, they intersect. Here, d is the distance between centers, and r1 and r2 are the radius of centers.



Case 1: d = r1 + r2



Case 2: r1 - r2 < d < r1 + r2

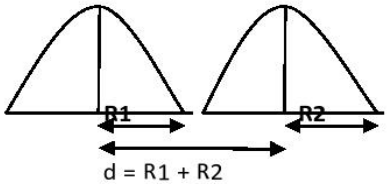


Fig 7 (a): d >= R1+R2

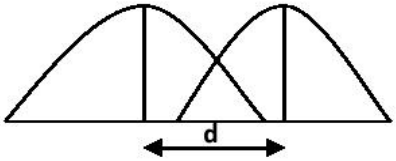


Fig 7 (b): d >= R1-R2

Similarly, in our approach, we considered the mean to be equivalent to the centre of the circle and the standard deviation to be equivalent to the radius of the circle. Suppose the distance between the mean is greater than or equal to the sum of the standard deviations. In that case, there is no overlap, as illustrated in Figure 7(a), and no normalised value is calculated. But if the distance between the mean is greater and equal to the difference in standard deviation, then, in this case, the classes are partially overlapped, as shown in Figure 7(b). Otherwise, the classes are completely overlapped, and we calculate the normalised value for these cases. For attribute x, the normalised value could be computed using the formula-

$$x_{\text{norm}} = \frac{a-b}{max-min} \times 100 \quad \text{----------(3)}$$

Where max and min are the total diameters of both classes, a and b are the common diameters in both classes. We will analyse each attribute separately and find their overlapped distance (common diameter) and normalised value. In accordance with their normalised value,

attributes are organised in ascending or descending order. Then the first k-best features are selected by adding one feature after each iteration of the forward or backward approach for the classification algorithm. Minimum k=3 features are selected together at the first iteration, and their accuracy has been noted at different k-fold values. Then the other features will be added one by one to the subset using the forward feature selection technique of the wrapper method. Then the average accuracy was calculated for each subset of features at different k-fold values. Ultimately, the K-best features with the highest accuracy will be selected.

Proposed Feature Selection Algorithm:

Step 1: Calculate mean and standard deviation for each attribute in every class.

Step 2: Assume mean of each attribute as a center and 3*standard deviation of each attribute as a radius for every class.

Step 3: If difference of mean of both class is greater than equal to summation of standard deviation of both class,

then

no overlap region

return 0

Step 4: If difference of means of both class is greater than equal to difference of standard deviation of both class,

then

partially overlap region, go to step 5.

else

completely overlap region, go to step 5.

Step 5: Calculate normalised value using equation (3) given above.

Step 6: Arrange the normalised value in ascending order.

Step 7: Apply k-fold validation on attributes.

Step 8: Calculate the average accuracy.

EXPERIMENTAL RESULTS

The dataset library for machine learning at UCI and Kaggle, two collections of databases used by data scientists and machine learning enthusiasts, are the main data sources used in this study.

From these publicly accessible services, we downloaded and used three datasets: two from Kaggle and one from UCI. The heart disease and diabetes databases are gathered from Kaggle, and the breast cancer dataset is taken from UCI. An open-source IDE called Jupyter Notebook has been used to implement the method and perform the experiments. This section discusses all of the tests and operations carried out on these datasets, as well as the results we obtained.

Datasets

This study uses publicly available data that was acquired from Kaggle and the UCI machine learning archive. The suggested system has been implemented on three distinct datasets taken from Kaggle and UCI [16, 18]. The first dataset, which is made up of 768 patient records and nine covariates, is for Pima Indian women who have diabetes and was taken from Kaggle [19]. The characteristics include the number of pregnancies, blood pressure, skin width, insulin, body weight, family history of diabetes, age, and results. The class attribute called outcomes indicates if a person has diabetes or not. The diabetes dataset had no missing values. Wisconsin Breast Cancer is the second dataset, and it has 699 instances with 11 attributes. It was retrieved from the UCI machine learning repository. Attributes are listed in the first 10 columns, while class attributes are presented in the 11^{th} column. These properties comprise sample code number, clump width, cell uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, empty nuclei, bland chromatin, normal nucleoli, and mitoses. Benign (non-cancerous) or Malignant (cancerous) are options for the class attribute. Malignant is denoted by four, while benign is represented by two. Since the ID numbers in the column were unnecessary and had no bearing on the suggested model's evaluation, we eliminated them. Some missing values in the Wisconsin Breast Cancer dataset have been filled in using the mean value for that particular column. The same dataset's class attribute originally had values of 2 and 4, which were changed to 0 and 1 accordingly to facilitate calculation [20]. The third dataset, the Cleveland Dataset (cardiovascular disease), has 303 samples and a total of 76 attributes. However, they are intimately associated with cardiac disease. The bulk of research only uses a maximum of 14 traits. They age, sex, maximal heart rate (thalach), exercise-induced angina (exang), ST depression (old peak), chest pain (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), slope, number of vessels (ca), heart rate (thal), and target. Target is a class attribute that contains 0 (absence) and 1 (presence). There were no missing entries in the heart disease dataset. All three datasets contain

two classes, "0" and "1," indicating whether there are any diseases present or not [21]. After data pre-processing, features and class attributes have been separated. Each attribute's mean and standard deviation have been computed and are classified into groups of 0 and 1.

The F1-score, recall, precision, and accuracy of all three datasets are displayed in Table 1. All these four are important concepts of machine learning. F1-score, recall, and precision are performance measures for pattern recognition and classifications in machine learning. The proportion of relevant occurrences among the collected instances is known as precision, whereas the percentage of pertinent examples that were really recovered is known as recall. The F1-score integrates a classifier's precision and recall into one metric by calculating their harmonic means.

Table 1:- All three dataset's accuracy, precision, recall, and F1 Score

DATASETS	PRECISION	RECALL	F1 SCORE	ACCURACY
HEART	75.75	89.28	81.96	81.36
DIABETES	57.4	72.09	63.91	76.74
BREAST	95.83	92	93.87	95.54

Tables 2, 3, and 4 show the importance of feature selection. All three described the procedure of best feature selection with the help of the k-fold validation method. Table 2 represents the best feature selection for diabetes. A minimum of three features are chosen and subjected to k-fold validation; the smallest value of k in the k-fold is 5, and the value changes in multiples of 5 (i.e., k = 5, 10, 15, 20, 25, and 30). Accuracy has been calculated based on these different k-fold values for the specific features. Similarly, for other features, accuracy has been calculated. After getting accuracy for all features on different values of k-fold validation, the average accuracy has been computed. This process has been applied to all three given datasets. Regarding the diabetes dataset, we got the highest average accuracy of 76.745% on six features: glucose, insulin, diabetes pedigree function, BMI, pregnancy, and blood pressure. These six features are the best in diabetes prediction. Figure 8 demonstrates the proposed healthcare system's examination of prediction accuracy for diabetes and existing relevant healthcare systems, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), K-Means, Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), J48, REP-TREE, and Multi-layer Perceptron (MLP). The proposed method's effectiveness is further verified by contrasting the suggested technique's performance with other published methods that

are also employed on the same dataset. As shown in Table 2, the suggested method performed better than previously published methods.

Table 2:-Diabetes k-best features

DIABETES DATASET							
K-BEST	GAUSSIAN NAÏVE BAYES ACCURACY						AVERAGE
FEATURES	5 FOLDS	10 FOLDS	15 FOLDS	20 FOLDS	25 FOLDS	30 FOLDS	ACCURACY
3	73.44	73.68	73.57	73.33	73.3	73.19	73.41833333
4	76.3	76.55	76.18	76.45	76.17	76.71	76.39333333
5	76.69	76.56	76.3	76.46	76.31	76.47	76.465
6	77.08	77.08	76.7	76.58	76.31	76.72	**76.745**
7	75.52	75.9	75.52	76.05	75.92	75.81	75.78666667
8	75.26	75.51	75.26	75.53	75.4	75.03	75.33166667

For heart disease, seven factors have a significant impact on the prediction. As shown in Table 3, the selected features are old peak, thal, exang, cp, thalach, ca, and sex, and achieve the highest accuracy of 81.36%. Figure 7 depicts a Cleveland Heart Disease comparison graph with 303 records and 14 characteristics. We compared the classification accuracy with five methods, which are C4.5, Fast Decision Tree (FDT), NB, DT, and KNN, but the results of the method we presented in this research work are slightly greater than all these methods.

Table 3: Heart Disease k-best features.

HEART DISEASE DATASET							
K-BEST	GAUSSIAN NAÏVE BAYES ACCURACY						AVERAGE
FEATURES	5 FOLDS	10 FOLDS	15 FOLDS	20 FOLDS	25 FOLDS	30 FOLDS	ACCURACY
3	69.26	73.92	73.22	73.91	74.28	74.24	73.13833333
4	74.87	77.5	77.84	78.85	79.2	79.87	78.02166667
5	76.2	78.18	78.17	78.54	78.87	78.84	78.13333333
6	77.2	78.81	79.46	81.08	80.46	80.75	79.62666667
7	80.5	81.17	80.8	81.79	81.82	82.12	**81.36666667**
8	78.51	80.48	81.49	81.81	81.82	82.15	81.04333333
9	78.85	81.16	81.25	81.52	82.17	81.54	81.08166667
10	78.85	80.48	80.82	80.5	81.82	81.18	80.60833333
11	77.85	80.48	80.15	80.83	80.82	81.48	80.26833333
12	77.85	80.48	79.49	80.16	80.51	80.81	79.88333333
13	77.2	80.47	79.15	80.14	80.48	80.81	79.70833333

Table 4 shows that all nine features are mandatory for breast cancer prediction in humans. The key characteristics for identifying breast cancer include thickness of the clumps, consistency

of cell size and shape, marginal adhesion, size of a single epithelial cell, bare nuclei, plain chromatin, regular nucleoli, and mitoses. Hence, 95.54% is the highest accuracy among all the average accuracy computed for k-best features.

Table 4:-Breast Cancer k-best features.

BREAST CANCER DATASET							
K-BEST	GAUSSIAN NAÏVE BAYES ACCURACY						AVERAGE
FEATURES	5 FOLDS	10 FOLDS	15 FOLDS	20 FOLDS	25 FOLDS	30 FOLDS	ACCURACY
3	90.84	91.13	91.15	91.13	91.13	91.15	91.08833333
4	93.71	93.56	93.58	93.71	93.7	93.86	93.68666667
5	94.42	94.71	94.44	94.56	94.56	94.43	94.52
6	95.56	95.71	95.58	95.85	95.85	95.86	95.735
7	95.28	95.42	95.43	95.42	95.42	95.44	95.40166667
8	95.56	95.56	95.58	95.56	95.56	95.58	95.56666667
9	95.56	95.42	95.58	95.56	95.56	95.58	**95.54333333**

Figure 10 demonstrates the visual depiction of the accuracy of the algorithm we used to predict breast cancer compared with the algorithms used in previous research. The classifiers we used for comparative analysis are REP-TREE, DT, NB, MLP, J48, SVM Polynomial Kernel, Feed Forward, Voted Perceptron, and Instance-Based Learner (IBK). The KNN algorithm is called IBK in WEKA software. The proposed algorithm has greatest accuracy when compared to the other available classifiers, as shown in a bar graph.

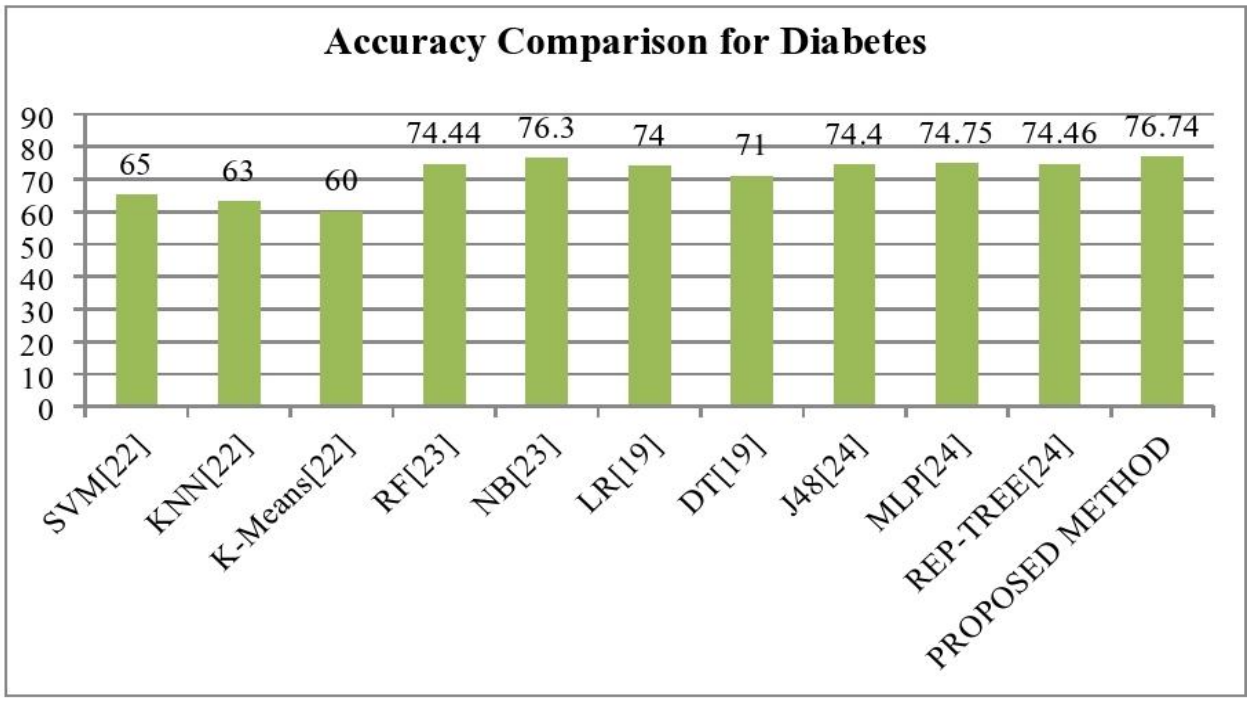


Fig. 8: Comparison Table for Pima Indian Diabetes Dataset (Kaggle)

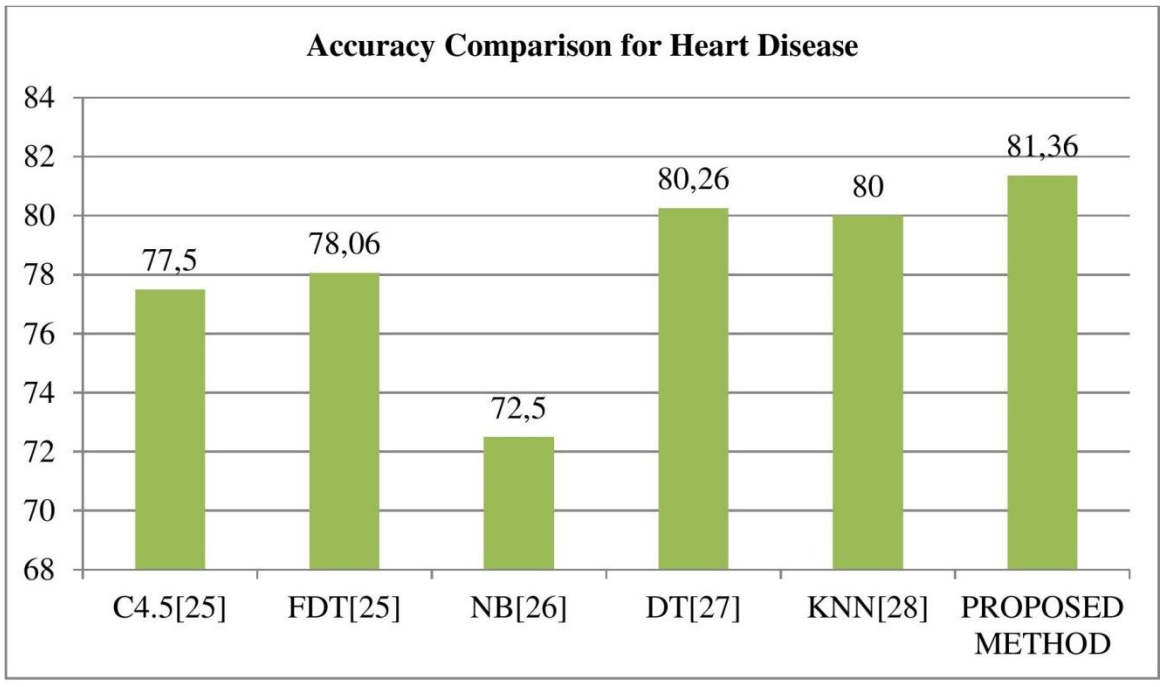


Fig. 9: Comparison Table for the Cleveland Heart Disease Dataset (Kaggle)

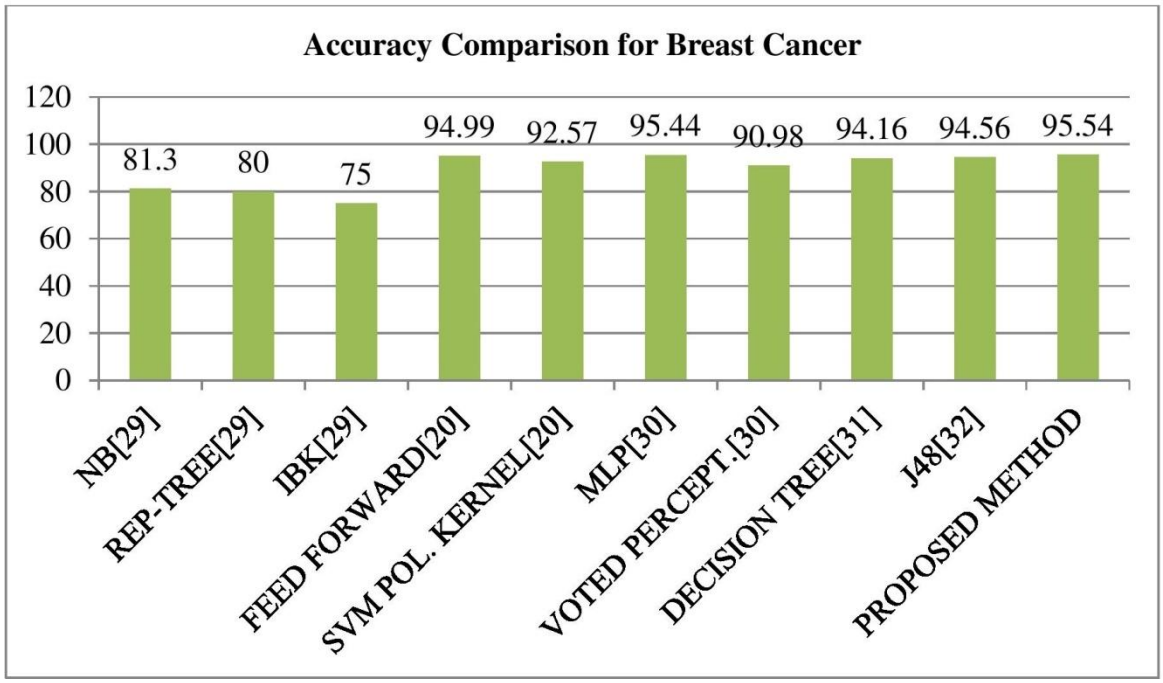


Fig. 10: Comparison Table for Wisconsin Breast Cancer Dataset (UCI)

CONCLUSIONS

This study proposes a feature selection method for continuous data relying on class overlapping. The concept of this approach is based on the intersection of two circles, which provides us with a clear view of whether the classes are overlapped. The same concept has been implied in our proposed method for finding the overlapped region when data is continuous. The prime goal of this approach is to select the K-best features for continuous data in order to make classification easier. The proposed algorithm will help in identifying three possible cases that come under "class overlapping," i.e., partial class overlapping, complete class overlapping, and no class overlapping, and calculate the normalised value of that region. Then, using forward feature selection, the Gaussian naive Bayes classifier, and the k-fold validation method, the k-best features were chosen. The model was implemented on three separate datasets, namely, diabetes, heart disease, and breast cancer, which are publicly available on the Kaggle and UCI repositories. We have contrasted the accuracy of the proposed approach with other machine learning methods that have been used for feature selection in previous studies. The suggested approach showed some improved results as compared with other methods.

For Pima Diabetes, the proposed approach gives a high accuracy of 76.74% in selecting 6 features, which are glucose, insulin, diabetes pedigree function, BMI, pregnancy, and blood pressure, using the suggested feature selection algorithm. There is an increase in accuracy as compared to the previously mentioned values shown in Table 5. For heart disease, we get 81.36% accuracy with this method, which is relatively better than the other past values, as shown in Table 6. This is the highest average accuracy among all cases on 7 features that include old peak, thal, exang, cp, thalach, ca, and sex. The Breast Cancer Dataset has shown 95.54% accuracy in selecting all 9 qualities, including clump thickness, homogeneity of cell size and shape, marginal adhesion, size of a single epithelial cell, naked nuclei, plain chromatin, common nucleoli, and mitoses shown in the above Table 7. In the future, this technique will help when dealing with continuous data for classification. The designed method is easy to implement, improves the efficiency of the system, and gives higher and more accurate results.

Acknowledgements:

The authors appreciate the support of the **Samrat Ashok Technological Institute (SATI), Vidisha.**

Conflict of Interest:

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

1. Cai, J., Luo, J., Wang, S. and Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing, 300, pp.70-79.
2. Uçar, M.K., 2020. Classification performance-based feature selection algorithm for machine learning: P-score. IRBM, 41(4), pp.229-239.
3. Lim, H. and Kim, D.W., 2021. Pairwise dependence-based unsupervised feature selection. Pattern Recognition, 111, p.107663.
4. Varma, P.R.K., Kumari, V.V. and Kumar, S.S., 2016. Feature selection using relative fuzzy entropy and ant colony optimization applied to real-time intrusion detection system. Procedia Computer Science, 85, pp.503-510.

5. Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A. and Bashir, K., 2019, January. Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623). IEEE.

6. Jain, D. and Singh, V., 2018. Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal, 19(3), pp.179-189.

7. Manzoor, I. and Kumar, N., 2017. A feature reduced intrusion detection system using ANN classifier. Expert Systems with Applications, 88, pp.249-257.

8. Verma, A.K., Pal, S. and Kumar, S., 2019. Comparison of skin disease prediction by feature selection using ensemble data mining techniques. Informatics in Medicine Unlocked, 16, p.100202.

9. Mishra, N.K. and Singh, P.K., 2020. FS-MLC: Feature selection for multi-label classification using clustering in feature space. Information Processing & Management, 57(4), p.102240.

10. Mirzaei, S., El Yacoubi, M., Garcia-Salicetti, S., Boudy, J., Kahindo, C., Cristancho-Lacroix, V., Kerhervé, H. and Rigaud, A.S., 2018. Two-stage feature selection of voice parameters for early Alzheimer's disease prediction. Irbm, 39(6), pp.430-435.

11. Magesh, G. and Swarnalatha, P., 2021. Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. Evolutionary intelligence, 14, pp.583-593.

12. Hasan, N. and Bao, Y., 2021. Comparing different feature selection algorithms for cardiovascular disease prediction. Health and Technology, 11, pp.49-62.

13. Sandhiya, S. and Palani, U., 2020. An effective disease prediction system using incremental feature selection and temporal convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 11(11), pp.5547-5560.

14. Nagarajan, S.M., Muthukumaran, V., Murugesan, R., Joseph, R.B., Meram, M. and Prathik, A., 2022. Innovative feature selection and classification model for heart disease prediction. Journal of Reliable Intelligent Environments, 8(4), pp.333-343.

15. Hasan, S.M.M., Mamun, M.A., Uddin, M.P. and Hossain, M.A., 2018, February. Comparative analysis of classification approaches for heart disease prediction. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

16. Kamel, H., Abdulah, D. and Al-Tuwaijari, J.M., 2019, June. Cancer classification using gaussian naive bayes algorithm. In 2019 International Engineering Conference (IEC) (pp. 165-170). IEEE.

17. Intersection of two circles - Higher Maths Revision. Retrieved from BBC. https://www.bbc.co.uk/bitesize/guides/z9pssbk/revision/4

18. Dubey, G.P. and Bhujade, R.K., 2021. Optimal feature selection for machine learning based intrusion detection system by exploiting attribute dependence. Materials Today: Proceedings, 47, pp.6325-6331.

19. Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A., 2018, September. Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

20. Showrov, M.I.H., Islam, M.T., Hossain, M.D. and Ahmed, M.S., 2019, December. Performance comparison of three classifiers for the classification of breast cancer dataset. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE.

21. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H. and Gutierrez, J., 2017, July. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE symposium on computers and communications (ISCC) (pp. 204-207). IEEE.

22. Lomte, R., Dagale, S., Bhosale, S., & Ghodake, S. 2018, Survey of different feature selection algorithms for diabetes mellitus prediction. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

23. Sivaranjani, S., Ananya, S., Aravinth, J. and Karthika, R., 2021, March. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 141-146). IEEE.

24. Verma, D. and Mishra, N., 2017, December. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In 2017 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 533-538). IEEE.

25. El-Bialy, R., Salamay, M.A., Karam, O.H. and Khalifa, M.E., 2015. Feature analysis of coronary artery heart disease data sets. Procedia Computer Science, 65, pp.459-468.

26. Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., 2013. Classification of heart disease using artificial neural network and feature subset selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence, 13(3), pp.4-8.

27. Ramalingam, V.V., Dandapath, A. and Raja, M.K., 2018. Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), pp.684-687.

28. Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. SN Computer Science, 1, pp.1-6.

29. Sakri, S.B., Rashid, N.B.A. and Zain, Z.M., 2018. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access, 6, pp.29637-29647.

30. Bayrak, E.A., Kırcı, P. and Ensari, T., 2019, April. Comparison of machine learning methods for breast cancer diagnosis. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). IEEE.

31. Marne, S., Churi, S. and Marne, M., 2020, March. Predicting breast cancer using effective classification with decision tree and k means clustering technique. In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 39-42). IEEE.

32. Mohammed, S.A., Darrab, S., Noaman, S.A. and Saake, G., 2020. Analysis of breast cancer detection using different machine learning techniques. In Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5 (pp. 108-117). Springer Singapore.