Hybridized Model using Clustering with Ensemble Classifier for Classification of Diseases

Rashmi Gupta
Ghasidas Vishwavidalaya (A Central University),
Bilaspur, 495001, India

Akhilesh Kumar Shrivas
Ghasidas Vishwavidalaya (A Central University),
Bilaspur, 495001, India

ABSTRACT

The ensemble technology in machine learning can be used for classification purposes and it is one of the most challenging tasks for the researchers who work for improving accuracy in disease predictions. This paper propose the hybrid model that is combination of Particle Swarm Optimization (PSO), K-means clustering and ensemble classifier for classification of five different types of medical diseases. The proposed hybrid model uses PSO technique for reducing the number of features from datasets, a k-means clustering approach for reducing the instances from dataset and ensemble classifiers is used to classify the different types of diseases from five different types of datasets. The ensemble classifier uses voting scheme to ensemble the individuals base classifiers like Multilayer Perceptron (MLP), Adaboost, Bagging, Fuzzy Unordered Rule of Induction Algorithm (FURIA), and Random Forest(RF) with Naive Bayes classified where Naive Bayes is considered as a meta learner. The experiment results reveal that the combination of the ensemble classifier with PSO and k-means algorithm is an efficient novel method for disease predictions. The improved accuracy was found in between 95% to 100%. The results are evaluated and also compared with different existing models which showed better performance than others.

Keywords

Particle swarm optimization (PSO), Machine learning, Ensemble classifier, Classification, Clustering, Multilayer Perceptron (MLP), Bagging, Adaboost, Random forest (RF), Voting, FURIA.

1. INTRODUCTION

A classification task in machine learning can be implementing using various single, hybrid, and ensemble algorithms for making predictions. It helps to classify each and every instance in the dataset into different groups using prior knowledge of the data. Now the researcher's mostly concerned and worked with the hybrid model and the ensemble learner system because it helps to predict data more accurately than single ones [5]. Ensemble learner algorithm combines the prediction results from multiple models and is designed to give a better performance. Both hybrid and ensemble models are based on the fusion concept, but slightly in different ways. Homogenous and weak models are used to combine in ensemble models, but heterogeneous model is used in the hybrid machine learning system. Therefore, both are used in real-world applications for solving different types of problems [4][9].

In this paper, we propose a novel hybrid model for various diseases predictions that uses the Particle Swarm Optimization (PSO) technique for obtaining the most relevant features from datasets, K-means clustering for reduced the unclustered instances or wrongly clusters instances from the datasets and voting ensemble classifier framework approach is used to ensemble the classifiers that classify the diseases of five different medical datasets.The proposed system introduces a novel approach by integrating the K-means clustering algorithm with Particle Swarm Optimization (PSO) and an ensemble classifier. This integration aims to reduce the number of instances in the dataset while improving the overall model performance.

Many researchers [17][23] had used both clustering and classification for enhancing the performance of classifiers in diagnosing and predicting diseases. Clustering is used to validate the data and used to remove the unclustered instances from the datasets and also effectively contributed to feature selection. K-means clustering is efficiently used to improve the quality of data and thus improved the performance of the system [18][19].

The rest section of this paper work is divided into four sections, where second section describes the related existing work that helps to develop the proposed concept, third section illustrate the overall methodology used in this work, fourth section explore the experimental work and results, and the last section conclude the proposed system and future implementation.

2. LITERATURE REVIEWS

Researchers followed various approaches to combine the prediction of various algorithms in ensemble learning to solve the particular problem. This section presented all the existing researchers who worked using ensemble learning and also a hybrid system for classification and feature selections methodology.

Leon, et al. [13] used voting ensemble classifiers with bagging and compared the performance in terms of accuracy. They also described different voting schemes and the result of this work reveals that a single transferable voting ensemble is a good alternative to plurality voting although it has some drawback of higher computational cost. Panthong & Srivihok [19] proposed a work on wrapper feature selection method based on the ensemble learning algorithm, and results showed that sequential feature selection (SFS) based on bagging algorithm using decision tree gave better results with an accuracy of 89.60%. Manonmani & Balakrishnan [16] proposed a framework that used ensemble feature selection technique with CKD datasets. The proposed experiment achieved higher prediction accuracy for deriving features. Thaseen et al. [25] proposed an ensemble-based approach that reduced the complexity of system and overcome the different issues in the existing ensemble-based intrusion detection system. This model utilized the classifier such as Support vector machine (SVM), Modified Naïve Bayes (MNB), LPBoost, and chi-square for feature selection in NSL-KDD datasets and gave 99% accuracy. Das et al. [8] constructed a model by using a neural network-based ensemble methodology and gained 89.01% classification accuracy, 80.95% sensitivity, and 95.91% specificity values for diagnosis the heart disease. To evaluate the performance of various ensemble methodologies, Das & Sengur [6] proposed a framework on three popular ensemble algorithms such as bagging, boosting, and random subspace and proved their efficiency over a single base classifier. Hambali et al. [11] proposed a model that used both homogenous and heterogeneous classifiers for breast cancer diagnosis. This experiment used ensemble methodology by using a decision tree algorithm, Naïve Bayes, and Synthetic Minority Oversampling Technique (SMOTE). The performance of Adaboost- random forest gave maximum performance than other ensemble methods. Zhang Y. et al.[29] proposed the framework using ensemble weighted voting classifier that combines the different classifiers. The system improved the performance with strong generalization ability. The results of classifiers combine according to the combination rule of weighted voting on differential solution. Cohagan, C.[4] conducted the experiment using voting ensemble approach that followed support, strength and democratic methods on 16 different datasets. They found that the democratic method is better than voting based on support, and also strength method. Leung & Stott Parker [14] concerned to develop an effective methodology in ensemble base voting learning. They also compared different voting methods in ensemble learning using bagging. Dietterich, T. G. [10] also performed comparisons for efficiency between 3 ensemble classifiers like bagging, boosting, and Randomization against C4.5 classifier. This system proved that boosting-C4.5 are given better results with no noise as compared to other combination, also Randomizing-C4.5 also gave improve result than Bagging-C4.5. Van Erp M. et al.[26] performed the discussion on unweighted voting method, confidence voting method and ranked voting method with two datasets for pattern recognition. Tigga N. P. et al. [30] have compared the performance of classifiers like logistic regression, K-NN,SVM, Naïve bayes, Decision tree and Random forest where Radom forest achieved highest accuracy as 94.10% for classification of Type 2 diabetics disease. Ahmed N. et al. [31] suggested different machine learning techniques with labeled and normalization preprocessing technique for classification of diabetic dataset with high accuracy. The proposed SVM with label-encoding and normalization improves the accuracy. This model also integrates in a web application using python flask web development framework. This experiment and results reveals that ML-based model with preprocessing technique classifies diabetes disease more accurately. Massari H.E. et al. [32] compared the different classifiers like SVM, KNN, ANN, Naive Bayes, Logistic regression, and Decision Tree with proposed Ontology classifiers for classification of diabetic disease. The proposed Ontology classifier achieved highest performance in terms of accuracy, precision, recall, F-measure and ROC area. Phongying M. and Hiriote S. [33] compared proposed model with four machine learning algorithms like decision trees, random forests, support vector machines, and K-nearest neighbors classifiers for classification of diabetic disease where proposed model achieved better performance in terms of accuracy precision, recall, f-score. Karthick K. et al. [34] compared the performance of SVM, Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest algorithm for classification of Cleveland heart disease (HD) dataset with chi-square feature selection technique . The random forest achieved highest 88.5% of accuracy with 13 features. Debal DA et al. [35] compared the performance of SVM ,DT and radom forest algorithm with feature selection technique for classification of chronic kidney disease. The experiment results show that random forest with recursive feature elimination achieved better performance than SVM and DT. AlyasIn T. et al. [36] suggested and compared various machine learning algorithms like decision tree, random forest algorithm, KNN, and artificial neural networks for classification of thyroid disease. The suggested random forest algorithm achieved better 94.8% accuracy and 91% specificity. Sun J. et al. [37] discussed the basic steps o skin lesion diagnosis based on various previous researches. This study found that various researchers have used machine learning based methods for classification of skin disease.

3. METHODS AND MATERIALS

Methods and material play very important role to develop the robust model. This section explores the dataset, and methodology used in this research work. This section also explores the flow of proposed system for classification of various medical diseases.

3.1 Datasets

This experiment has been designed with the five most known datasets taken from the UCI machine learning repository [38]. Table1 described the details of all datasets used in this research work. All these medical datasets contain different features, classes, and characteristics need to be preprocessing before using any algorithm.

3.2 System flow of the proposed method

The proposed system followed the overall process in four steps (i) feature selection process using PSO algorithm, (ii) Reducing the wrongly or unclustered instances using K-means clustering, (iii) Ensemble classifiers using voting ensemble approach (iv) Comparing and evaluating the proposed system with the available system. Figure 1 represents the flow diagram of our proposed system. In this proposed system, the PSO feature optimization technique applied to reduce the features set, after that k-means clustering technique applied to reduce the instances of datasets. Then, ensemble the different classifiers like MLP, Random forest, FURIA, Bagging and Adaboost with Naïve bayes classifier using voting ensemble technique. The optimized medical datasets are applied to different machine learning techniques and ensemble models, and compared their performance in terms of accuracy. The proposed hybrid ensemble model predicts the various diseases correctly for giving better solution.

3.3 Preprocessing and feature reduction

The feature selection method is used to retrieve the most relevant feature from the datasets and it is essential to improve the classification performance for the system. It plays a very important role in the classification process to reduce the unknown data and for developing an accurate diagnosis system [10][22].

Table 1. Description of the medical datasets

Dataset	Number of class	Number of attributes	Number of instances	characteristics
Heart disease	2	14	1025	Categorical, real
CKD	2	25	400	real
Thyroid	4	30	3772	Categorical, real
Diabetes	2	9	768	Categorical, integer
Dermatology	5	35	366	Categorical, integer

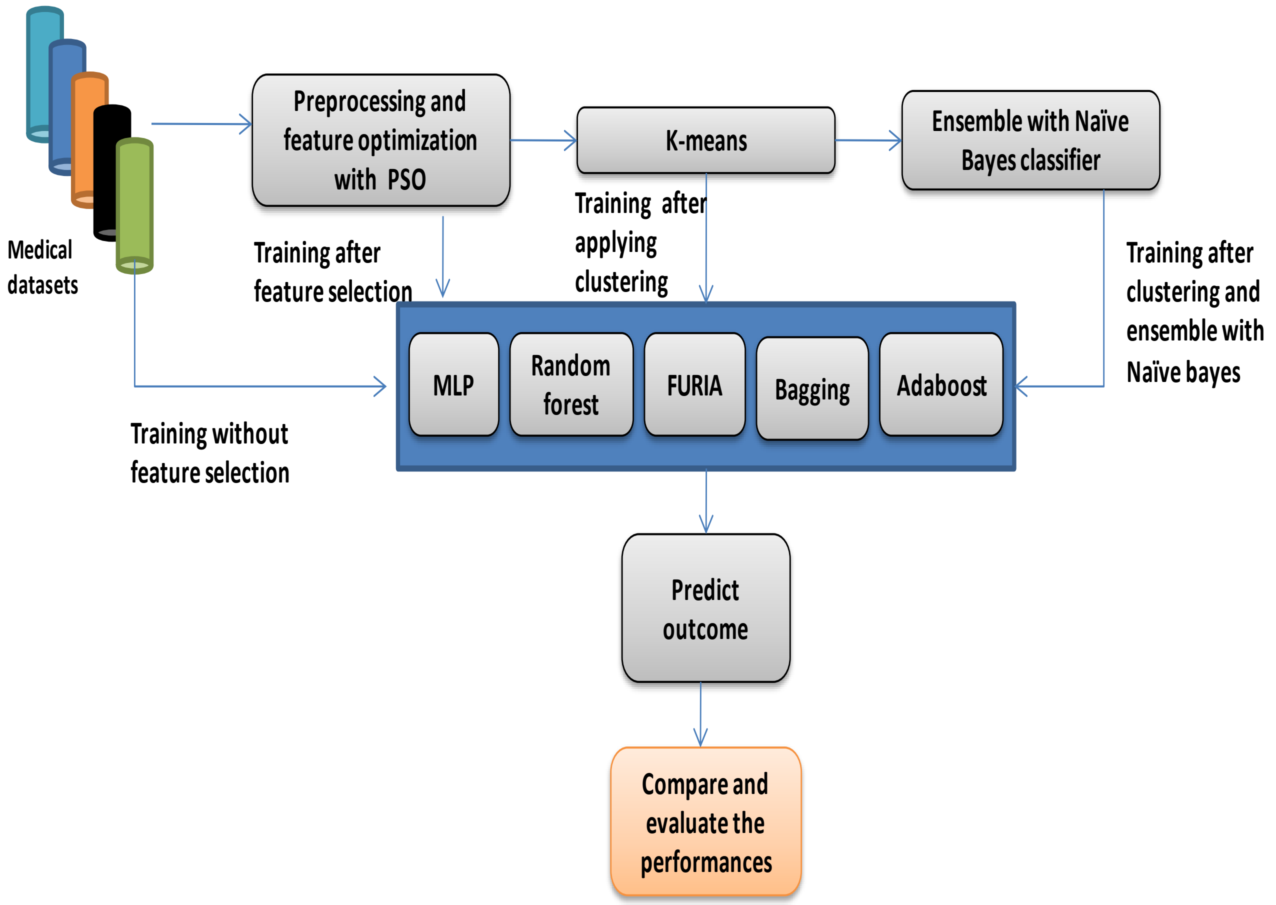
Medical datasets
Preprocessing and feature optimization with PSO
K-means
Ensemble with Naïve Bayes classifier
Training after feature selection
Training after applying clustering
Training after clustering and ensemble with Naïve bayes
MLP
Random forest
FURIA
Bagging
Adaboost
Training without feature selection
Predict outcome
Compare and evaluate the performances

Figure 1. Architecture diagram of the proposed hybrid ensemble system

In our proposed hybrid system, PSO is used to select the most optimal required attributes. This multiple optimization methods decreases the complexity of the inference classification algorithm during testing and provides more effective and rapid operation for disease prediction, thus improve the accuracy of the system [20].

PSO is a stochastic global search optimization algorithm based on flocks of birds looking for resources, moving around a search space at a specific velocity for searching better solution [3][12]. This facilitates classification algorithms by reducing data complexity [14][24].

3.4 K-means clustering

Clustering is an unsupervised machine learning algorithm used to partition the data into n-clusters. This technique is used in data analysis to classify the data and to find the optimal solution for all data points. In this paper, K-means is a clustering algorithm concerns for reducing the unclustered instances from datasets prior to the classification or ensemble learning process. This makes improving the performance of the classifiers and hence increases the accuracy of the proposed system [8][15].

3.5 Ensemble algorithm

The ensemble learning techniques in machine learning combine different models and algorithms to solve a particular problem. It also refers to multiple prediction models and had proved its effectiveness over the last few years [4]. Generally ensemble members are constructed in two ways- the first one is using single learning algorithm, and the second one is using different learning algorithm [29]. The second method is the good for proving the efficiency and compares the performance against different algorithms. Using this approach we can combine the prediction of different algorithm and work for reducing the error in predictions of result. In this work, a voting ensemble methods applied to ensemble classifiers like Adaboost, bagging, MLP (Multilayer Perceptron), Fuzzy Unordered Rule of Induction Algorithm (FURIA), and Random forest as a base learner with Naïve Bayes considered as a Meta learner classifier. The main objective behind a ensemble classifier with k-means clustering is to improve the performance of various diseases [1][11]. Voting ensemble technique is used to summing the prediction made by different algorithms, or it goes the average prediction made by regression models [2]. Using this technique, the proposed system achieved higher prediction accuracy. Depending on the choice of base classifiers, the performance of an ensemble method can outperform a single classifier [5]. Therefore researcher's main concern for implementing this method is the selection of base classifier and train the base classifier for finding the optimal weight. This help to ensemble the classifiers based on weighted voting.

4. EXPERIMENTAL WORK AND RESULTS

This section describes the experimental work of proposed system and implementation of models for classification, clustering, and hybrid ensemble technique for diagnosing and classification of various diseases.

The ensemble of machine learning based classification models provide a useful means of averaging error that comes from individual classifiers and help in reducing the generalization error for classification [14]. This evaluation tests, how effectively the data are used with k-means and ensemble for improving the performance of the system with five different medical datasets. The implementation begins with the preprocessing and feature selection process which can be done using PSO. This optimization process helped in selecting the most discriminatory features for classification. K-means clustering algorithm helps to find the most usable instances and gave powerful handling data for classification [8]. The primary focus of this system is to give strength to the classifier and reduce the deficiencies of the single algorithm by using clustering and ensemble classification model and thus able to produce a better system with improved accuracy. Therefore, ensemble the Naïve Bayes with different classifiers after performed the feature optimization and clustering to reduce features and instances. Naïve Bayes is one of the simplest and common learning algorithms. A vote ensemble approach is used to ensemble the two or more similar types of trained classifiers to improve the performance as compared to individuals classifiers. This research work has used voting ensemble technique to ensemble various individual classifiers like Adaboost, bagging, MLP, FURIA, and random forest as a base learner with Naïve Bayes as a Meta learner. Training of the base classifiers and ensemble classifier can be done using 10 –fold cross validation method. The main idea behind is to build a robust hybrid ensemble model that is combination of PSO, K-means and ensemble model for improving the classification accuracy with five different types of medical diseases. WEKA machine learning tool kit is used to develop the proposed hybrid ensemble mode for classification of various diseases. For better evaluation and interpretation, calculated and compared the training accuracy of algorithm at each and every stage of proposed flow model. Tables 3 illustrate the accuracy of proposed model with diabetes dataset, Table 4 illustrates the accuracy of proposed model with thyroid dataset, Table 5 illustrates the accuracy of proposed model with dermatology dataset, Table 6 illustrates the accuracy of proposed model with heart disease dataset, and Table 7 illustrates the accuracy of proposed model with CKD dataset. It could be observed that comparisons of the classification accuracy gave a deeper looker at how the algorithms are being used. This method strongly impacts classification results and it also takes less execution time and reduces uncertainty in their decision but the computational cost seems to be high.

The ensemble of machine learning based classification models provide a useful means of averaging error that comes from individual classifiers and help in reducing the generalization error for classification [14]. The test results how effectively when the data are used with k-means and ensemble for improving the performance of the system with five different medical datasets. The implementation begins with the preprocessing and feature selection process which can be done using PSO. This optimization process helped in selecting the most discriminatory features for classification. K-means clustering algorithm helps to find the most usable instances and gave powerful handling data for classification [8]. The primary focus of this system is to give strength to the classifier and reduce the deficiencies of the single algorithm by using clustering and ensemble classification model and thus able to produce a better system with improved accuracy. Therefore, it significantly improves classification performance by optimizing both instance selection and feature selection. The proposed approach effectively reduces dataset size while maintaining or even enhancing classification accuracy. The ensemble classifier ensures robust and reliable predictions by leveraging the strengths of multiple base learners, reducing overfitting, and improving generalization. It ensemble the Naïve Bayes with different classifiers after performed the feature optimization and clustering also reduce features and instances. Naïve Bayes is one of the simplest and common learning algorithms.

Table 2. Descriptions of the medical datasets with optimization technique

Dataset	Reduced features of dataset after PSO	Reduced instances of dataset K-means	Reduced dataset after PSO and k-means
Heart disease	10	729	10x729
CKD	18	122	18x122
Thyroid	20	1808	20x1808
Diabetes	4	235	4x235
Dermatology	17	95	17x95

Table 3. Classification accuracy of proposed model with diabetes dataset

classification algorithms	**Accuracy in %**					
	Original data	**Reduced data using PSO**	**Reduced data using PSO and k-means**	**Ensemble with Naïve Bayes in original data**	**Ensemble with Naïve Bayes in optimized data using PSO**	**Ensemble with Naïve Bayes in optimized data using PSO and k-means (proposed model)**
MLP	75.39	75.52	99.62	76.82	76.39	**99.24**
FURIA	74.47	75.52	98.49	76.17	76.3	**98.49**
RF	75.78	74.73	99.24	76.95	76.56	**98.68**
Adaboost	74.34	74.34	98.12	77.08	76.82	**98.87**
bagging	75.78	74.47	98.31	76.95	77.21	**98.49**

Table 4. Classification accuracy of proposed model with thyroid dataset

Classification algorithms	**Accuracy in %**					
	Original data	**Reduced data using PSO**	**Reduced data using PSO and k-means**	**Ensemble with Naïve Bayes in original data**	**Ensemble with Naïve Bayes in optimized data using PSO**	**Ensemble with Naïve Bayes in optimized data using PSO and k-means (proposed model)**
MLP	94.69	96.23	97.91	95.36	95.04	**99.33**
FURIA	99.52	97.61	99.94	99.46	97.5	**99.84**
RF	99.39	96.92	99.94	96.87	95.7	**99.64**
Adaboost	95.38	95.38	100	95.57	94.61	**99.94**
Bagging	99.57	97.4	100	98.93	95.14	**99.74**

Table 5. Classification accuracy of proposed model with dermatology dataset

Classification algorithms	**Accuracy in %**					
	Original data	**Reduced data using PSO**	**Reduced data using PSO and k-means**	**Ensemble with Naïve Bayes in original data**	**Ensemble with Naïve Bayes in optimized data using PSO**	**Ensemble with Naïve Bayes in optimized data using PSO and k-means (proposed model)**
MLP	98.36	96.72	100	98.08	97.81	**99.63**
FURIA	93.98	95.62	98.52	97.81	97.81	**99.63**
RF	97.26	96.72	99.63	97.54	98..36	**99.63**
Adaboost	50.27	50.27	50.18	97.54	98.63	**99.53**
bagging	95.62	95.62	97.41	97.81	98.63	**98.52**

Table 6. Classification accuracy of proposed model with heart disease dataset

Classification algorithms	Accuracy in %					
	Original data	Reduced data using PSO	Reduced data using PSO and k-means	Ensemble with Naïve Bayes in original data	Ensemble with Naïve Bayes in optimized data using PSO	Ensemble with Naïve Bayes in optimized data using PSO and k-means (proposed model)
MLP	95.51	91.41	100	93.56	90.04	**100**
FURIA	100	99.51	100	100	99.6	**100**
RF	100	100	100	83.12	93.26	**100**
Adaboost	84.29	83.12	100	86.24	84.78	**100**
bagging	94.55	95.12	100	89.75	88.68	**100**

Table 7. Classification accuracy of proposed model with CKD dataset

Classification algorithms	Accuracy in %					
	original data	Reduced data using PSO	Reduced data using PSO and k-means	Ensemble with Naïve Bayes in original data	Ensemble with Naïve Bayes in optimized data using PSO	Ensemble with Naïve Bayes in optimized data using PSO and k-means (proposed hybrid ensemble model)
MLP	83.33	99.25	100	98.5	98.75	**100**
FURIA	97.5	96.5	99.28	97.75	97	**98.92**
RF	99	98.75	100	96.25	96.5	**99.64**
Adaboost	96.25	96	100	97	96.75	**100**
Bagging	97.25	97.25	100	97	97.25	**100**

5. RESULT ANALYSIS AND COMPARISON

The result analysis can be performed by comparing the accuracy for different combinations with proposed hybrid ensemble classifiers for deeper evaluation and interpreting the proposed system. However, the proposed method was tested on various benchmark datasets to evaluate its impact on classification performance shown in Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6. It depicts that the comparisons are made using base learning classifiers in original data, datasets after optimizing features using PSO, data after PSO–K-Means clustering, Ensemble with Naïve Bayes in original data, Ensemble with Naïve Bayes in optimized data using PSO, Ensemble with Naïve Bayes in optimized data using PSO and K-means. This proved that when proposed hybrid ensemble model classified data and improved the performance in terms of accuracy reached up to 100%. This result reveals that the combination of the ensemble model with K-means is a good approach for improving the performance and efficiency of the classification model. Finally**,** the proposed K-means + PSO + Ensemble system demonstrates significant improvements across multiple datasets with higher accuracy **(**over traditional models**)**, reduced misclassification errors (False Positives & False Negatives), faster training time and better generalization**.** Finally our suggested model achieved better accuracy as compared to other models discussed in literature.

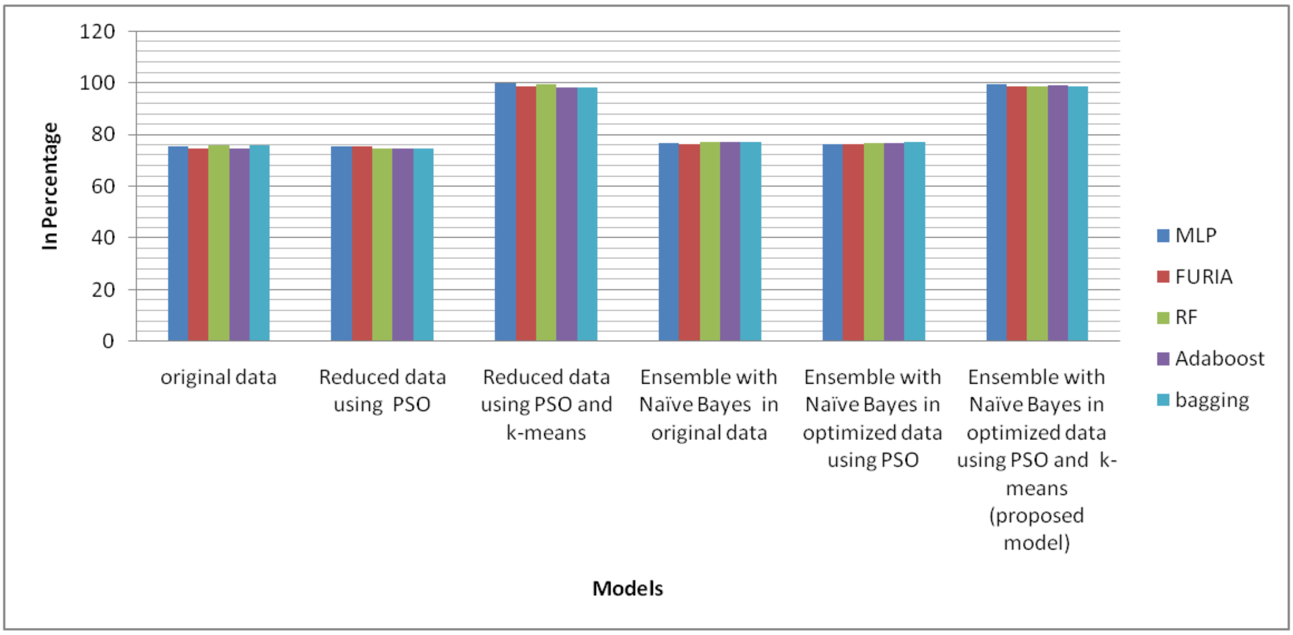

Figure 2. Accuracy chart of diabetes disease dataset

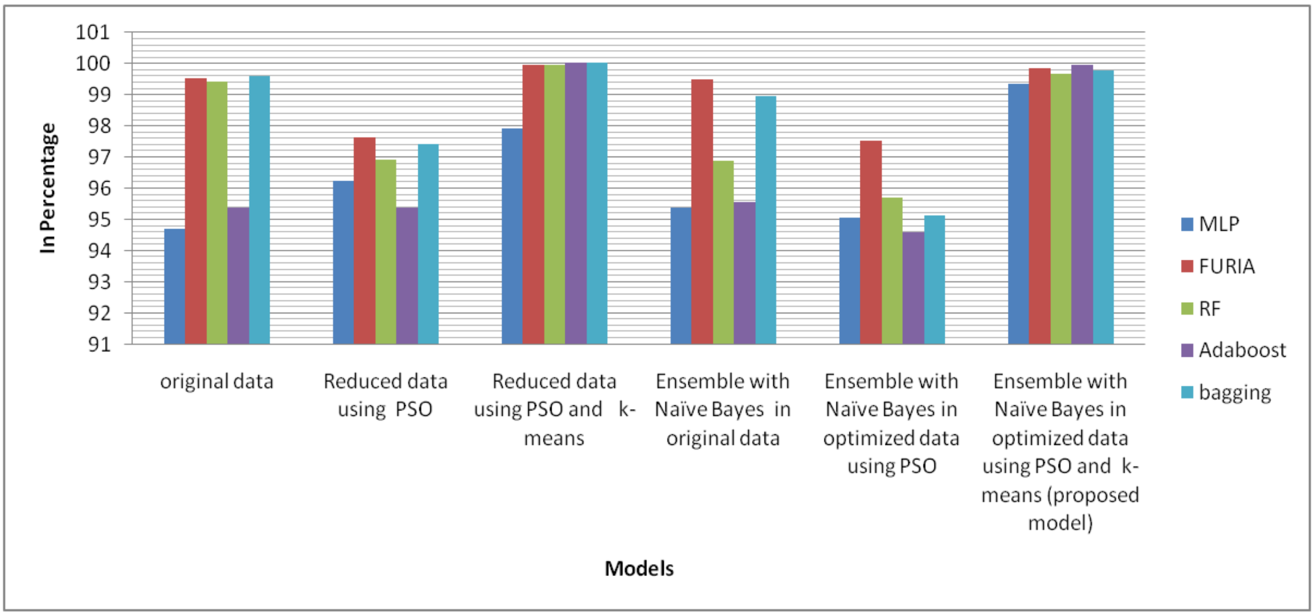

Figure 3. Accuracy chart of Thyroid disease dataset

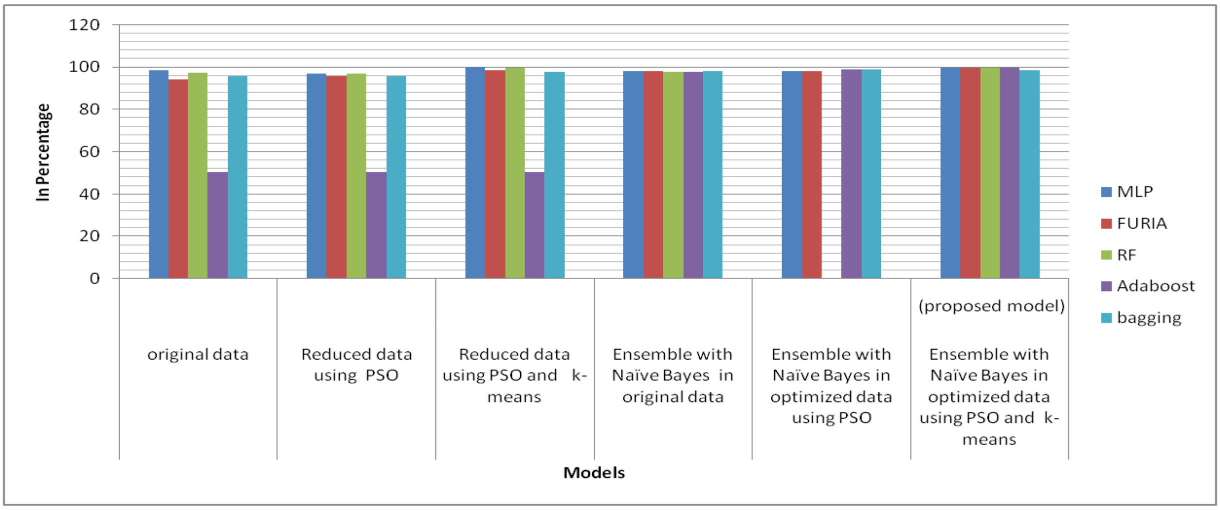

Figure 4. Accuracy chart of dermatology disease dataset

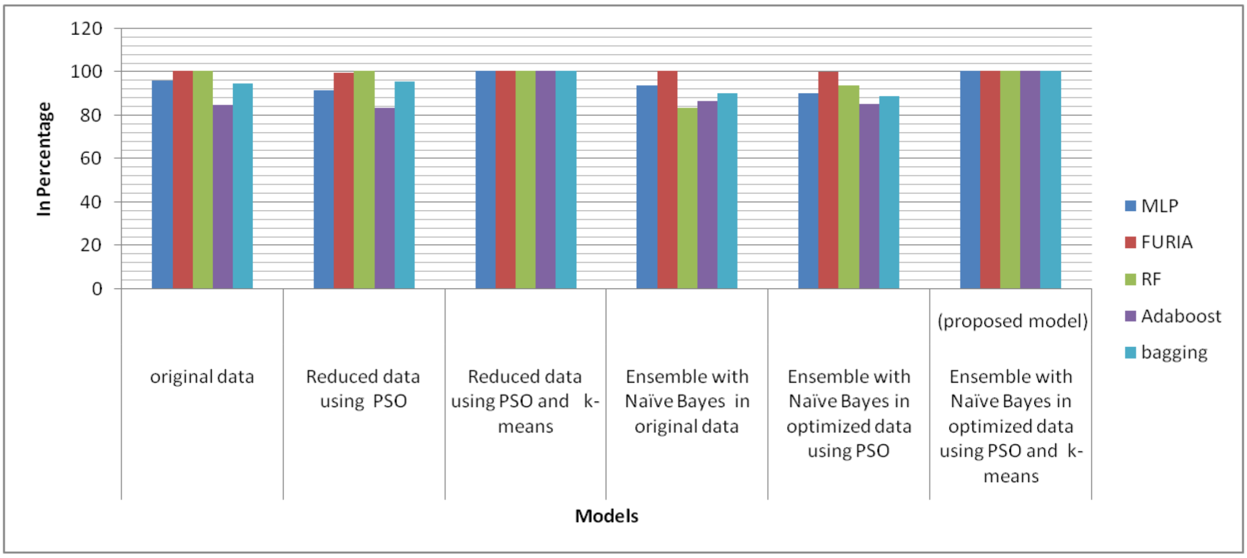

Figure 5. Accuracy chart of heart disease dataset

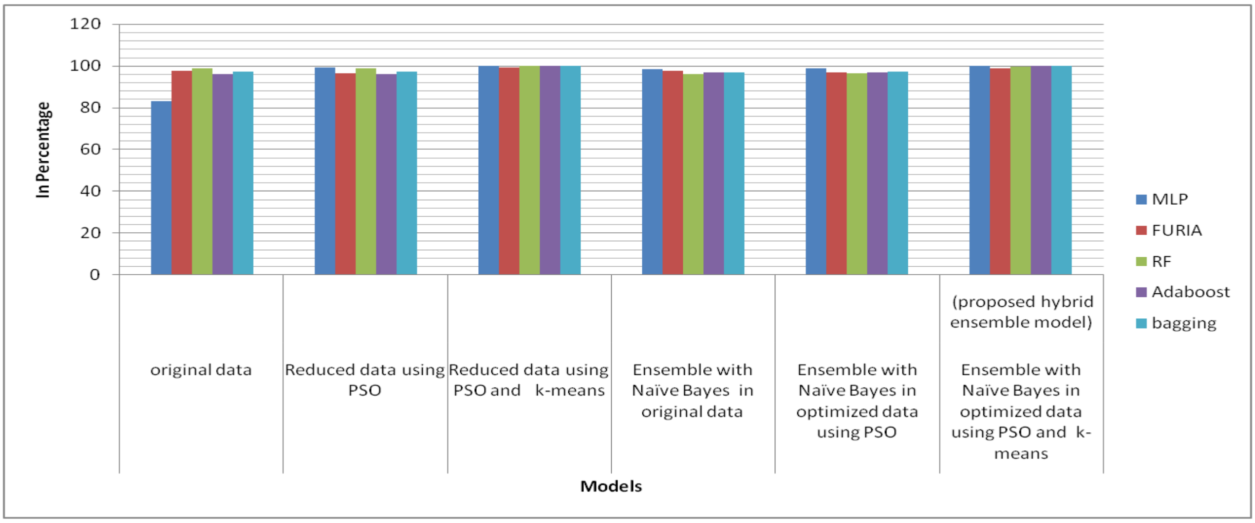

Figure 6. Accuracy chart of chronic kidney disease dataset

6. CONCLUSIONS

In this proposed novel work, the ensemble hybrid method with k-means clustering is used in five different medical datasets before performing the classification algorithms which gave potential improvement and better performance in accuracy. Comparisons with different strategies and combinations had also been done which showed that the proposed method gave a better alternative approach for disease diagnosis. It also suggested that the k-means clustering algorithm with hybrid ensemble effectively contributed and strongly impact in classification results. In evaluation, an ensemble classifier is employed to enhance classification performance by combining predictions from multiple base learners. This ensures robustness and reduces the risk of overfitting. A reduced dataset and optimized features give lower computational complexity and ensemble model ensures reliable predictions on unseen data. It also test how effectively the data are used with this hybrid ensemble technology produced better and more accurate prediction results and taking less computation time also. This framework provides effective decision support system and helpful in further implementation.

In future, Implement this experiment with different clustering approach and voting ensemble classifier framework for more validation and also do more work for reducing computational cost.

7. REFERENCES

[1] Alaba A, Maitanmi S, Ajayi O. "An Ensemble of classification techniques for Intrusion Detection Systems", International Journal of Computer Science and Information Security, vol. 17, no. 11, pp. 24-33, 2019.

[2] Bauer E, Kohavi R., "Empirical comparison of voting classification algorithms: bagging, boosting, and variants", Machine Learning, vol. 36, no.1, pp. 105–139, 1999.

[3] Carvalho M, Ludermir TB., "Hybrid training of feed-forward neural networks with particle swarm optimization", Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics. 4233 LNCS, pp. 1061–1070, 2006.

[4] Cohagan C, Grzymala-Busse JW, Hippe ZS., "A Comparison of Three Voting Methods for Bagging with the MLEM2 Algorithm", Proceedings of the 11th international conference on Intelligent data engineering and automated learning, pp. 118-125, 2010.

[5] Cordón O, Kazienko P, Trawiński B., "Hybrid and ensemble methods in machine learning", New Generation Computing, vol. 29, no. 3, pp. 241–244, 2011.

[6] Das R, Sengur A., "Evaluation of ensemble methods for diagnosis of valvular heart disease", Expert Systems with Applications, vol. 37, no. 7, pp. 5110–5115, 2010.

[7] Das R, Turkoglu I, Sengur A., "Diagnosis of valvular heart disease through neural networks ensembles", Comput Methods Programs Biomed, vol. 3, pp. 185–191, 2008.

[8] Das R, Turkoglu I, Sengur A., Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, vol. 36, no. 4, pp. 7675–7680, 2009.

[9] Das S, Abraham A, Konar A., "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Xplore, vol. 38, no. 1, pp. 218–237, 2008.

[10] Dietterich T G. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization", Kluwer Academic Publishers, Manufactured in The Netherlands, Machine Learning, vol. 40, pp. 139–157, 2000.

[11] Hambali MA, Saheed YK, Oladele TO, Gbolagade M D. "Adaboost Ensemble Algorithms for Breast Cancer Classification", International Journal of Advances in Computer Research Quarterly, vol. 10 , no. 2, pp. 31-52, 2019.

[12] Kazemi Y, & Mirroshandel SA, "A novel method for predicting kidney stone type using ensemble learning", Artificial Intelligence in Medicine, vol. 84, pp. 117–126, 2018.

[13] Leon F, Floria SA, Badica C. "Evaluating the effect of voting methods on ensemble-based classification", Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, pp. 1–6, 2017.

[14] Leung KT, Stott PD. "Empirical Comparisons of Various Voting Methods in Bagging KDD '03", Proceedings of the ninth ACM SIGKDD International conference on Knowledge discovery and data mining, pp. 595-600, 2003.

[15] Lin KC, Hsieh YH. "Classification of Medical Datasets Using SVMs with Hybrid Evolutionary Algorithms Based on Endocrine-Based Particle Swarm Optimization and Artificial Bee Colony Algorithms", Journal of Medical Systems, vol. 39, no. 10, pp 1-9, 2015.

[16] Manonmani M, Balakrishnan S. "An ensemble feature selection method for prediction of chronic diseases", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 5, pp. 7405–7410, 2020.

[17] Mohebian MR, Marateb HR, Mansourian M, Mañanas MA, Mokarian F., "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning", Computational and Structural Biotechnology Journal, vol. 15, pp. 75–85, 2017.

[18] Naaz E, Sharma D, Sirisha D, Venkatesan M. , "Enhanced K-means clustering approach for health care analysis using clinical documents", International Journal of Pharmaceutical and Clinical Research, vol. 8, no. 1, pp. 60–64, 2016.

[19] Panthong R, Srivihok A., "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm", Procedia Computer Science, vol. 72, pp. 162–169, 2015.

[20] Patil BM, Joshi RC, Toshniwal D., "Hybrid prediction model for Type-2 diabetic patients", Expert Systems with Applications, vol. 37, no. 12, pp. 8102–8108, 2010.

[21] Patil D, Agrawal B, Andhalkar S, Biyani R, Gund M, Wadhai DVM., "An Adaptive parameter free data mining approach for healthcare application", International Journal of Advanced Computer Science and Applications, vol. 3, no. 1, pp. 55–59, 2012.

[22] Purwar A, Singh SK. , "Hybrid prediction model with missing value imputation for medical data", Expert Systems with Applications, vol. 42, no. 13, pp. 5621–5631, 2015.

[23] Santos V, Datia N, Pato MPM., "Ensemble Feature Ranking Applied to Medical Data", Procedia Technology, vol. 17, pp.223–230, 2014.

[24] Sasikala S, Appavu Alias Balamurugan S, Geetha S., Multi Filtration Feature Selection (MFFS) to improve

discriminatory ability in clinical data set, Applied Computing and Informatics, vol. 12, no. 2, pp. 117–127, 2016.

[25] Thaseen IS, Kumar CA, Ahmad A., "Integrated Intrusion Detection Model Using Chi-Square Feature Selection and Ensemble of Classifiers", Arabian Journal for Science and Engineering, vol. 44, no. 4, pp. 3357–3368, 2019.

[26] Van EM, Vuurpijl L, Schomaker L. "An overview and comparison of voting methods for pattern recognition", Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR, pp. 195–200, 2002.

[27] Verma L, Srivastava S, Negi PC. "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", Journal of Medical Systems, vol. 40, no. 7, pp. 1-7, 2016.

[28] Xue B, Zhang M, Member S, Browne WN, "Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach", IEEE Transactions on Cybernetics,1–16, 2012.

[29] Zhang Y, Zhang H, Cai J, Yang B. , “A weighted voting classifier based on differential evolution”, Abstract and Applied Analysis, 1-6, 2014.

[30] Tigga NP, Garg S. “Prediction of Type 2 Diabetes using Machine Learning Classification Methods”, Procedia Computer Science, vol. 167, pp. 706–716, 2020.

[31] Ahmed N, Ahammed R, Islam MM, Uddin MA Akhter A, Talukder MA , Paul BK., “ Machine learning based diabetes prediction and development of smart web application”, International Journal of Cognitive Computing in Engineering, vol. 2, pp. 229–241, 2021.

[32] Massari HE, Sabouri Z, Mhammedi S ,Gherabi N., “Diabetes Prediction Using Machine Learning Algorithms and Ontology”, Journal of ICT Standardization, vol. 10, no. 2, pp. 319–338, 2022.

[33] Phongying M, Hiriote S., “Diabetes Classification Using Machine Learning Techniques”, Computation, vol. 11, pp. 1-17, 2023.

[34] Karthick K, Aruna SK, Samikannu, Kuppusamy R, Teekaraman Y, Thelkar AR., “Implementation of a Heart Disease Risk Prediction Model Using Machine Learning”, Computational and Mathematical Methods in Medicine, pp. 1-14, 2023.

[35] Debal, DA, Sitote, TM., “Chronic kidney disease prediction using machine learning techniques”, J Big Data, vol. 9, no. 109, pp. 1-19, 2022.

[36] Alyas T, Hamid M, Alissa K, Faiz T, Tabassum N, Ahmad A., “Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach”, BioMed Research International, pp. 1-10, 2022.

[37] Sun J, Yao K, Huang G, Zhang C, Leach M, Huang K, Yang X., “Machine Learning Methods in Skin Disease Recognition: A Systematic Review”, Processes.2023, vol.11, no. 4, pp. 1-16, 2023.

[38] Open UCI repository: https://archive.ics.uci.edu/datasets (Accessing date: 25-10-2022).