Improving Prediction Accuracy Using Hybrid Machine Learning Algorithm on Medical Datasets

Dr. Anitha Avula V, Arba Asha

Abstract— In Computer Aided Decision(CAD) systems, machine learning algorithms are adopted to assist a physician to diagnose disease of a patient. The purpose of this study is to improve the prediction accuracy on medical datasets by hybridizing machine learning algorithms. In this paper Hybrid Machine Learning algorithm is proposed using two supervised algorithms, Naïve Bayes and JRIP. The methodology adopted in this paper for proposing new Hybrid Machine Learning Algorithm is implemented by using R programming language and weka software tool. Further, comparative study is made between individual algorithms and proposed hybrid algorithm to prove the improvement in prediction accuracy on medical datasets. The proposed algorithm shows enhanced performance compared to the individual classifiers and assist the physician in diagnosis.

Index Terms— Hybrid algorithm, Machine learning, prediction accuracy, supervised algorithms

1 Introduction

Machine learning has evolved from the computational learning theory and pattern recognition. It is the most effective method used in the field of data analytics in order to predict something by devising some models and algorithms[1]. Machine learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions[2]. There are many machine learning algorithms typically grouped by either learning style(i.e. supervised learning, usupervised learning, semi-supervised learning) or by similarity in form or function (i.e. classification, regression, decision tree, clustering, deep learning, etc.).. Currently, the application of machine learning in medical diagnosis is a new trend for large medical data applications. Most of the diagnosis techniques in medical field are systematized as intelligent data classification approach. In Computer Aided Decision (CAD) systems, information technology methods are adopted to assist a physician to diagnose disease of a patient. Aside from other traditional classification problems, medical dataset classification problems are also applied in the future diagnosis. Generally, patients or doctors are not completely informed about the cause (classification result) of the disease, but also will be made known of their symptoms that drive the cause of disease, which is the most important of their medical dataset classification problem.[3]. Many authors built different hybridized algorithms. In this study a new approach is used to propose the new Hybrid Machine Learning Algorithm which will be useful for the physician to diagnos the disease of a patient with an efficient predication accuracy. Further to proceed and implement the proposed hybrid algorithm, related work on medical datasets is discussed in section 2. The methodology used for building proposed hybrid machine learning algorithm using R Programming and Weka tool is discussed in Section 3. Section 4 discsusses comparative study and the Section 5, includes conclusion and recommendation for future work.

- *Dr. Anitha. V, Department of Computer Science, Faculty of Informatics, Havassa University, Hawassa, Ethiopia, PH+251937215423., E-mail:gamya21@gmail.com India PH:+919849444087, +919849446033*
- *Arba Asha, is currently pursuing masters degree program in computer science in Hawassa University, Hawassa, Ethiopia, PH-+251916542544 E-mail: arba40asha@gmail.com*

2 Related Work

In this section related work done on medical datasets using hybrid machine learning algorithms is discussed. The researchers strived to improve the accuracy by using different combinations of machine learning algorithms to build the hybrid algorithm.

The performance of individual classifiers can be enhanced by using the hybridization method[4]. It is an important and latest area of research as compared to individual learning approaches [5]. Hybrid and ensemble methods in machine learning have attracted a great attention of the scientific community over the last years [6]**.** Both ensemble models and hybrid methods make use of the information fusion concept but in slightly different way. In case of ensemble classifiers, multiple but homogeneous, weak models are combined [7], typically at the level of their individual output, using various merging methods, which can be grouped into fixed (e.g., majority voting), and trained combiners (e.g., decision templates) [8]. Hybrid methods, in turn, combine completely different, heterogeneous machine learning approaches [9]. In literature, there are different ways of classifying the training/ test instances into one of the predefined categories, like (1) Individual models, (2) Hybrid models and (3) Ensemble based models [10]. Individual approach involves using a single statistical or machine learning technique for classification. The hybrid and ensemble models are efficient and robust because they combine the complementary features of more than one learning technique and overcome the weakness of individual techniques. The hybrid models can be stand alone, transformational, tightlycoupled or fully coupled [11]. As per [12]hybrid models are of 4 types: Classification combined with Classification, Classification combined with Clustering, Clustering combined with Clustering and Clustering combined with

Classification. Ensemble learning uses various base classifiers combined using a particular strategy of combination such as bagging, boosting, voting, etc.

Abhishek and Kumar in [13] developed a hybrid classifier algorithm by merging Decision Tree and Naïve Bayes algorithms which will classify the Fitness data set. The classification accuracy of the Hybrid Classifier has enhanced by 15.79 % and 3.6 % as compared to DecisionTree and Naïve Bayes classifier.

In [14] authors designed a general hybrid adaptive ensemble learning framework (HAEL), and apply it to address the limitations of random subspace-based classifier ensemble approaches (RSCE). The experiments on the real-world datasets from the KEEL dataset repository for the classification task and the cancer gene expression profiles showed that: 1) HAEL works well on both the real-world KEEL datasets and the cancer gene expressionprofiles and 2) it outperforms most of the state-of-the-artclassifier ensemble approaches on 28 out of 36 KEEL datasets and 6 out of 6 cancer datasets.

Tharaha and Rashika in [15]concluded that the performance level of the hybrid algorithm (Decisiontree and Artificial neural network) is better than that of the individual performance of the algorithms. Artificial neural network has the highest performance when compared with Decision tree algorithm. In addition, they found that the large datasets can easily be trained and tested in using these algorithms to predict the diseases that are expected according to the datasets.

In [16], a new system was proposed for breast cancer classification. The new system uses a hybrid of K-means and Support Vector Machine (SVM). The proposed algorithm was compared with different classifier algorithms. The experimental results showed the effectiveness of the proposed algorithm and how it can obtain better results.

In [17] researcher proposed using k Nearest neighbor algorithm (kNN) and Naïve Bayes with imputation techniques which was used instead of removing the values that are missing from the Mammographic Mass data. The system was evaluated using different performance criteria such as accuracy, sensitivity, and specificity and ROC analysis.

Ramana et al. [18]develop a classification model to predict liver disease diagnosis using five popular classification algorithms and evaluate the performance of each model in termsof accuracy, precision, sensitivity and specificity. The study showed that the performances of all classifiers are better in one dataset (AP Liver dataset) as opposed to the other (BUPA Liver dataset) due to highly significant attributes such as total countof bilirubin, direct bilirubin and indirect bilirubin in the APdataset

A hybrid algorithm was presented [19] to combine the cAnt-Miner2 and the mRMR feature selection algorithms. The proposed algorithm was experimentally compared to cAnt-Miner2, using some public medical data sets to demonstrate its functioning. The experiments were very promising and the proposed approach is better in terms of accuracy, simplicity and computational costthan the original cAnt-Miner2 algorithm.

Another study in [20] demonstrated that the effectiveness of an unsupervised learning technique which is k-means clustering in improving supervised learning technique which is naïve bayes.The results showed that integrating k-means clustering with naïve bayes with different initial centroid selection could enhance the naïve bayes accuracy in diagnosing heart disease patients.

In [21] authors hybrid the genetic algorithm and the k-nearest neighbor algorithm in order to design efficient classifier model for breast cancer classification and they achieved high classification performance.

3 METHODOLOGY

This section contains the methodology used for proposing the new hybrid algorithm.

Step-1: Data Collection:

The source of data for proposed algorithm is University of California(UCI) repository for machine learning and the Weka dataset.

.

Step-2: Developing Proposed Hybrid Machine Learning Algorithm

The major activities undertaken in this phase are:

- Identifying the requirements of hybrid algorithm to be developed.
- Proposing the methodology used for developing the algorithm and implementing using machine learning tools i.e. R programming language and Weka software tool.
- Evaluating the results obtained by hybrid algorithm**.**

Proposed Hybrid Machine Learning Algorithm

1. Input dataset D
2. Preprocess the dataset
   - ✓ replace missing values
3. Apply Jrip algorithm on the preprocessed dataset
   - ✓ Jrip generates a number of rules
4. Apply Naïve Bayes algorithm on each rule generated from Jrip algorithm.
5. Remove misclassified instances then apply again Naïve Bayes algorithm

6. Take the average of the accuracy
7. Output: Hybrid algorithm

The above proposed hybrid algorithm was implemented using two machine learning tools. They are Weka 3.9.2 software and R programming language.
Weka is a workbench for machine learning which is intended to aid in the application of machine learning techniques to a variety of real-world problems. Waikato Environment for Knowledge and Algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. R is an open source programming language and software environment that supports statistical computing and graphics

3.1 Implementation using R Programming Language

The approach used for implementing hybrid algorithm using R programming is the output of each Jrip rule will become the input for the Naïve Bayes algorithm.

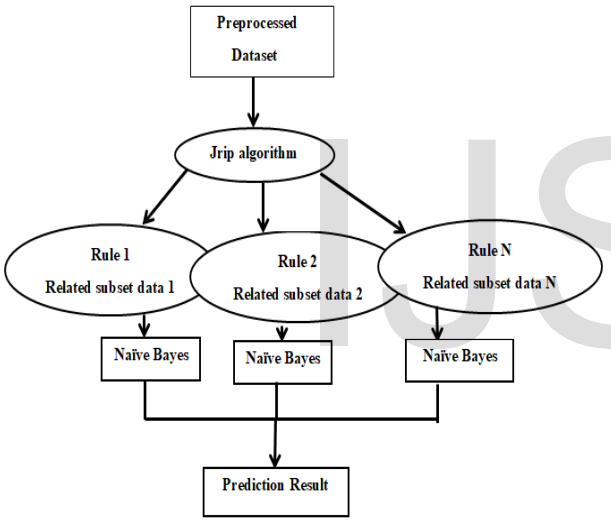

Fig 1. Flowchart for implementing proposed algorithm

The advantage of the proposed algorithm over the existing hybrid is that our algorithm will extract the rules for the input dataset using Jrip and each rule will be satisfied by a subset of dataset. On each such subset, the Naïve Bayes algorithm is implemented whose output will give more accuracy to predict the result of test dataset.
RWeka incorporates R interface classes for each key "group" of functionality delivered by Weka and to be interfaced (currently, classifiers, clusters, associators, filters, loaders, savers, and stemmers).
By importing data to the Weka software, explorer environment does the pre-processing work on the dataset. Preprocessing includes the replacement of missing values. This preprocessed dataset is imported and proposed hybrid algorithm is implemented using R programming. To support Jrip classifier, the RWeka package is installed in R. (install.Packages ("RWeka")).
The preprocessed dataset was given initially to the Jrip algorithm in R software environment. The algorithm generates a number of rules based on the given dataset. The JRip rule example based on the diabetes dataset is given below.

```
JRIP rules:
===========

(plas >= 155) => class=tested_positive (122.0/24.0)
(age >= 31) and (plas >= 115) and (mass >= 30.1) => class=tested_positive (91.0/31.0)
 => class=tested_negative (555.0/110.0)

Number of Rules : 3
```

Fig 2. JRIP rule generated by proposed algorithm for breast cancer dataset

From Fig 2, first rule is interpreted by taking the"plas" attribute, if it is greater or equal to the "155" then the class of instance is "tested_posetive". "122" is the number of instances that are classified as "tested_posetive" and "24" is the number of misclassified instances out of "122".
Second rule is interpreted by taking the attribute "age" If it is greater than or equal to "31", attribute "plas" is greater than or equal to "115" and attribute "mass" is greater than or equal to "30.1" then the class of instance is "tested_posetive".
Third rule will be considered when first and second rule fails and classified as "tested_negative".

3.2 Implementation using Weka Tool

The algorithms of our hybrid technique i.e. Jrip and Naïve Bayes are ensembled by using Weka software tool. The study is done by dividing the dataset using if-else java source code. According to the number of rules generated from the Jrip algorithm, the Naïve Bayes classifier works on each rule separately. The check for the accuracy is done by removing the misclassified instances and better result were found. Finally, the average of the each rule's accuracy was taken. Misclassified instances are removed to increase classification [22], [23]. Weka software tool is used to remove misclassified instances(pima_diabetesweka.filters.unsupervised.instance.RemoveMisclassified, Wweka.classifiers.bayes.NaiveBayes-C-1-F0-T0.1-I2).

4 Results

This study set out to enhance the prediction accuracy of hybrid machine learning algorithm. Hence, this section shows the analysis results and discussion about individual and proposed hybrid machine learning classifiers prediction accuracy on selected medical dataset. Here we have selected two supervised machine learning algorithms, they are Jrip algorithm and Naïve Bayes algorithm. By combining two selected algo-

rithms, the proposed hybrid algorithm was produced. Whose methodology of implementation is already discussed in section 3. The following results shown in the graphs are the outcomes of both ways of implementing our proposed hybrid i.e, using R programming language and using Weka software tool. While using Weka software tool, average probabilities combination rule of ensembling is done , as it gives better prediction performance on selected dataset.

4.1 Experiment on breast cancer dataset

A) Implementation using Weka ensembling

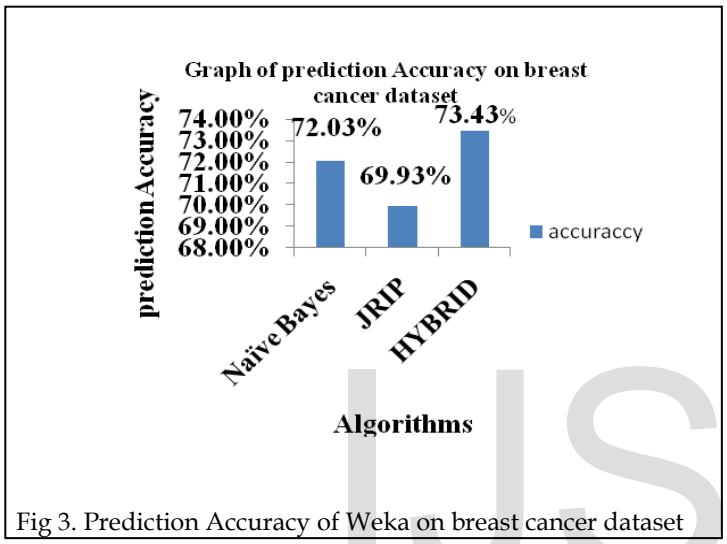

Fig 3. Prediction Accuracy of Weka on breast cancer dataset

The above graph in Fig 3, shows that the performance of selected individual and proposed hybrid classifier using breast cancer dataset to predict accuracy using Weka ensembling technique. It can be observed that prediction accuracy outperforms over individual classifiers.

B) Implementation using R Programming

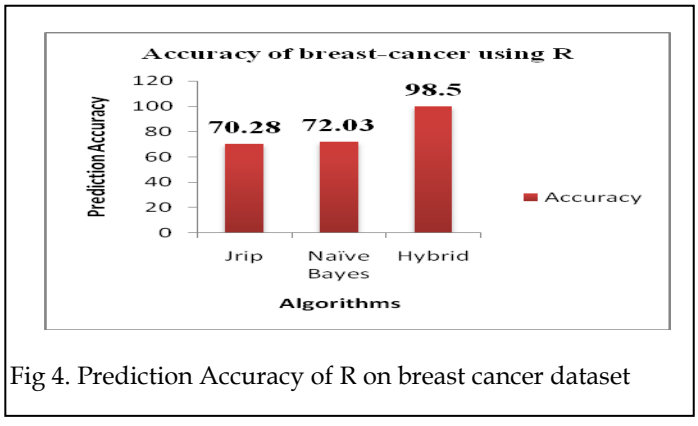

Fig 4. Prediction Accuracy of R on breast cancer dataset

The above graph in Fig 4, shows the prediction accuracy of Jrip, Naïve Bayes algorithm and proposed hybrid algorithm on breast-cancer dataset using R programming language. The prediction accuracy of the proposed hybrid algorithm is better than the other two individual classifiers after removing the misclassified instances.

4.2 Experiment on diabetes dataset

A) Implementation using Weka ensembling techniques

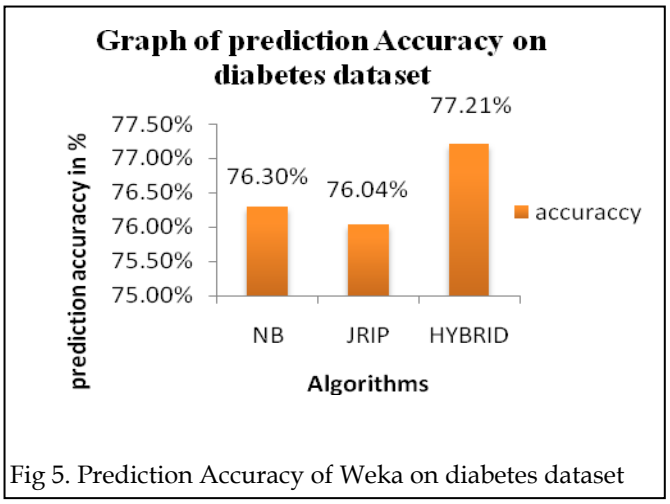

Fig 5. Prediction Accuracy of Weka on diabetes dataset

The above graph in Fig 5, shows the prediction accuracy of individual classifiers and hybrid classifiers on diabetes dataset using Weka ensembling technique. In this case, it can be observed the improvement in prediction accuracy when proposed hybrid algorithm is used.

B) Implementation using R Programming

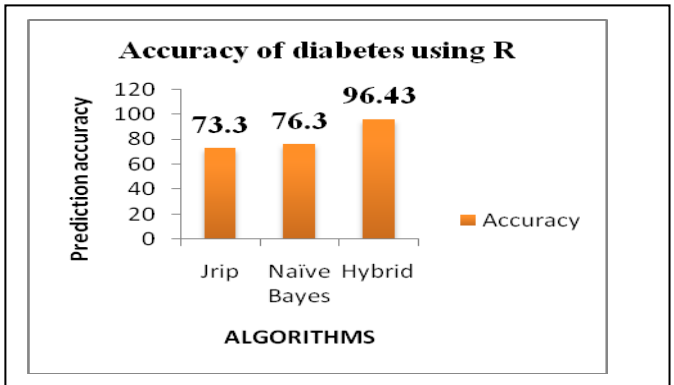

Fig 6. Prediction Accuracy of R on diabetes dataset

The above graph in Fig 6, shows the prediction accuracy of selected individual algorithms and proposed hybrid algorithm on diabetes dataset using R programming language. The proposed hybrid algorithm has better prediction accuracy than the individual algorithms.

The following table Table 1, shows the comparison of prediction accuracy of Jrip, Naïve Bayes algorithm and the proposed Hybrid algorithm using ensemble technique of

Weka on few datasets.

S. Nº	Dataset	Naïve bayes	Jrip	Hybrid
1.	Breast cancer	72.03%	69.93 %	**73.43%**
2.	Liver disorder	55.36%	64.64 %	**66.38%**
3.	Pima-Diabetes	76.30%	76.04 %	**77.21%**
4.	Hepatitis	83.87%	77.42 %	**84.52%**
5.	Chronic_Kidney_Disease	94.50%	96.00 %	**97.25%**
6.	Cardiotocography	96.75 %	100 %	**100%**
7.	Hypothyroid	95.23 %	99.47 %	**99.04%**
8.	cleveland-14-heart-disease	82.83 %	81.85 %	**83.83 %**
9.	heart-statlog	83.7037 %	78.89 %	**84.07 %**
10.	Dermatology	97.27 %	86.88 %	**94.53 %**

Table 1: Comparison Accuracy of Proposed Hybrid Algorithm using Weka ensemble technique

The overall results show higher values of accuracy for most datasets when we compare the individual and hybrid algorithm according. It can be observed that in many cases, the hybrid algorithm outperforms over the individual classifiers.

Proposed hybrid algorithm using R programming is tested on various datasets such as diabetes, breast cancer and liver-disorder dataset. The results of which are shown in the following tables i.e. Table 2, Table 3, Table 4.

Dataset	Generated rules	Accuracy before removing MCI	Accuracy after removing MCI
Diabetes	Rule-1	79.51%	93.88%
	Rule-2	65.93%	100%
	Rule-3	80.18 %	95.42%
Average		**75.20%**	**96.43%**

Table 2: Prediction accuracy on diabetes dataset

The above Table 2, shows the prediction accuracy that is obtained by applying proposed Hybrid algorithm in which Naïve Bayes algorithm is used on each rules generated from the Jrip algorithm. The prediction accuracy obtained after removing misclassified instance is better than the accuracy before removing the misclassified instances.

Dataset	Generated Rules	Accuracy before removing MCI	Accuracy after removing MCI
Breast cancer	Rule-1	66.66%	99%
	Rule -2	45.45%	98%
	Rule -3	73.03%	98%
Average		**61.71%**	**98.5%**

Table 3: Prediction Accuracy on breast_cancer dataset

The Table 3 shows the prediction accuracy of breast cancer dataset. It demonstrates the prediction accuracy before and after the removing of misclassified instances (MCI).

Dataset	Generated rules	Accuracy before removing MCI	Accuracy after removing MCI
Liver disorder	Rule-1	74.12%	99%
	Rule-2	65.71%	99%
	Rule-3	73.77%	99%
Average		**71.2%**	**99%**

Table 4:Prediction Accuracy on liver-disorder dataset

Table 4 shows prediction accuracy on liver disorder dataset.

S. Nº	Dataset	Proposed Hybrid using Ensemble Accuracy	Proposed Hybrid using R Accuracy
1	Pima diabetes	77.21%	96.43%
2	Breast-cancer	73.43%	98.5%
3	Liver-disorder	66.38%	99%

Table 5: Comparison of proposed ensemble and proposed Hybrid algorithm using R

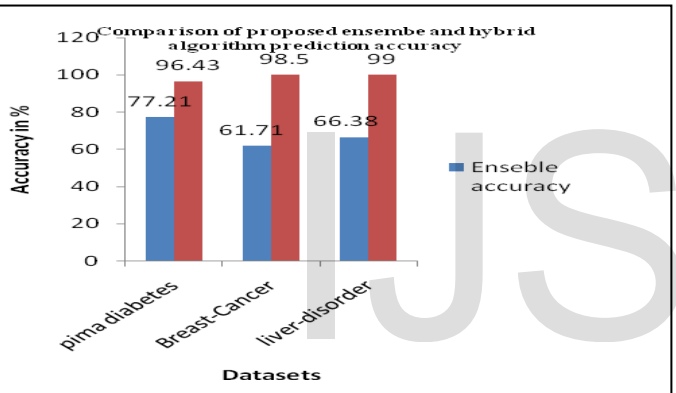

Fig 7: Comparison of Proposed Hybrid Algorithm using Weka Ensembling and R

The above Table 5 and graph of Fig 7, shows the comparison of proposed hybrid algorithm using ensemble technique of Weka and R programming on few datasets. It can be observed that hybrid algorithm implemented using R have more prediction accuracy compared to ensemble technique.

5 Conclusion

Machine learning systems make medical professionals faster and smarter in their diagnosis. As a result, it reduces uncertainty in their decisions, thereby reducing costs,risks and saving valuable time. In this study, the proposed ensemble and hybrid algorithm demonstrate that hybrid machine-learning techniques perform better than the individual algorithms on selected medical datasets. The proposed hybrid algorithm composed of Jrip and Naïve Bayes algorithm. To implement the proposed hybrid algorithm Weka and R machine learning tools were used. Selected individual algorithms separately and proposed hybrid algorithm is applied on different datasets using average probability combination rule of weka. 10-cross validation test option is used to get better prediction accuracy. The proposed hybrid algorithm using R outperform over the proposed ensemble algorithm plus the selected individual algorithms. The result shows that the hybrid machine-learning algorithm is the key to improve the prediction accuracy of individual machine learning algorithms.

6 Recommendation

In this study, some medical datasets are used to check prediction accuracy of algorithms, so it is better to use various category datasets that have different size for further work. Initially Jrip algorithm is used and then Naïve Bayes is used which means order of evaluation of selected algorithm is static. There is no possibility of selecting naïve bayes algorithm as the first algorithm and then jrip as the next depending on the type of datasets. If we can have the choice of dynamically selecting the order of the algorithms to work in the hybrid technique, the prediction accuracy can be further increased. Hence, dynamic implementation of proposed hybrid algorithm is possible and can be taken as the future research work.

References

[1] S. Angra and S. Ahuja, "Machine learning and its applications: A review", 2017.
[2] Nvidia; Stanford;McKinsey & Co., "What is Machine Learning?", [Online]. Available: https://www.techemergence.com/what-is-machine-learning/. [Accessed: 05-Oct-2018].
[3] C. V. S. and S. N. Deepa, "Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier", Sci. World J., vol. 2016, pp. 1–1, 2016.
[4] S. Dahiya, "Credit Modelling using Hybrid Machine Learning Technique," pp. 103–106, 2015.
[5] D. K. M. R. Malhotra, "Evaluating consumer loans using neural networks," Omega –The Int. J. Manag. Sci., pp. 83–96, 2003.
[6] E. Lughofer, "Hybrid and Ensemble Methods in Machine Learning J . UCS Special Issue," vol. 19, no. 4, pp. 457–461, 2013.
[7] K. Kajdanowicz, T., Kazienko, P., "Boosting algorithm with sequence-loss cost function for structured prediction," Springer, pp. 573–580, 2010.
[8] L. Kuncheva, "Combining pattern classifiers: Methods and algorithms", 2004.
[9] W. Castillo, O., Melin, P., Pedrycz, "Hybrid Intelligent Systems: Analysis and Design (Studies in Fuzziness and Soft Computing)," Springer, pp. 55–64, 2007.
[10] N. P. S. S. Dahiya, S.S Handa, "Credit Evaluation using Ensemble of various classifiers on reduced feature set", unpublished.
[11] A. R. G. A. Bahrammirzaee, "Hybrid Credit ranking intelligent system using expert system and artificial neural networks", Appl. Intell., vol. 34, pp. 28–46, 2011.
[12] C.F. Tsai, M.L. Chen, "Credit Rating by Hybrid Machine Learning Techniques", Appl. Soft Comput., vol. 10, pp. 374–380, 2010.
[13] L. Col and A. Lal, "Hybrid Classifier for Increasing Accuracy of Fitness Data Set," pp. 1246–1249, 2017.
[14] Z. Yu, S. Member, L. Li, J. Liu, and G. Han, "Hybrid Adaptive Classifier Ensemble", IEEE Trans. Cybern., pp. 1–14, 2014.
[15] T. S. K, "Hybrid Artificial Neural network and Decision Tree algorithm for Disease Recognition and Prediction in Human Blood Cells", 2017.
[16] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-

2 diabetic patients," Expert Syst. Appl., vol. 37, no. 12, pp. 8102–8108, 2010.

[17] C. Güzel and F. Engineering, "Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation," AWER Procedia Inf. Technol. Comput. Sci., vol. 4, pp. 401–407, 2013.

[18] M. S. P. B. and N. B. V. B. V. Ramana, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," Int. J. Database Manag. Syst., vol. 3, no. 2, pp. 101–114, 2011.

[19] I. Michelakos, E. Papageorgiou, and M. Vasilakopoulos, "A hybrid classification algorithm evaluated on medical data," Proc. Work. Enabling Technol. Infrastruct. Collab. Enterp. WETICE, pp. 98–103, 2010.

[20] M. Shouman, A. Defence, F. Academy, A. Defence, F. Academy, and R. Stocker, "i ntegrating naive bayes and k-means clustering with different initial centroid selection methods in the diagnosis," no. August 2014, p. 125–137, 2012.

[21] B. M. Abed et al., "A hybrid classification algorithm approach for breast cancer diagnosis", 2016 IEEE Ind. Electron. Appl. Conf., pp. 269–274, 2016.

[22] M. R. Smith and T. Martinez, "An Extensive Evaluation of Filtering Misclassified Instances in Supervised Classification Tasks," pp. 1–29, 2013.

[23] M. R. Smith and T. Martinez, "Improving classification accuracy by identifying and removing instances that should be misclassified," Proc. Int. Jt. Conf. Neural Networks, no. September 2011, pp. 2690–2697, 2011.