International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Anemia Selection in Pregnant Women by using Random prediction (Rp) Classification Algorithm

Dithy M.D, V KrishnaPriya

Abstract: ***Anemia is the global hematological disorder that occurs in pregnancy. The feature selection of unknown logical knowledge from the large dataset is capable with data mining techniques. The paper evaluates anemia features classes of Non-anemic, Mild and Severe or moderate in real time large-dimensional dataset. In the previous works, Anemia diseases can be classified in a selection of approaches, based on the Artificial Neural Networks (ANN), Gausnominal Classification and VectNeighbour classification. In these previous studies attains the proper feature selection with classification accuracy but it takes large time to predict the feature selection. So the current paper to overcome the feature selection, computational time process presents an improved Median vector feature selection (IMVFS) algorithm and new RandomPrediction (RP) classification algorithm to predict the anemia disease classes (Mild, Not anemic and Severe and moderate) based on the data mining algorithms. The results have shown that the performance of the novel method is effective compared with our previous Classification of ANN, Gausnominal and VectNeighbour classification algorithms. As the Experimental results show that proposed RandomPrediction (RP) classification with (IMVFS) feature selection methods clearly outperform than our previous methods.***

Index Terms: ***Anemia, Data mining, Random forest, Median Vector, Feature Selection***

I. INTRODUCTION

Data mining is a method to distinguish and substitute raw data into meaningful knowledge information, is gradually more being used in a combination of fields like business intelligence, marketing, scientific discoveries, biotechnology, multimedia and Internet searches. Data mining is an interdisciplinary field merging concepts from machine learning, statistics and natural language processing.

Anemia is required functioning red blood cells (RBCs) that guides to need of oxygen-carrying capability, sourcing abnormal complications throughout lifetime [1]-[2]. Through pregnancy, there is an uneven boost in plasma level, up to 50 percent, RBC 33 percent, and Hb 18-20 percent mass. In adding there is noticeable command for additional irons throughout pregnancy particularly in the next half of pregnancy. So, physiological anemia is due to the shared effect of hemodilution & negative iron balance.

In Anemia disease categorization, there is a quantity of data mining algorithms and tools obtainable to reflect on data mining tool Weka [3] to review on the data. Along with classification methods take a collection of classified samples (training set) and employ it for training the algorithms. With the qualified algorithms, prediction of the test sample data finds a place based on the representations and policies are extracted from the training set. The paper used this strategy for improved feature selection algorithms and random prediction classification of anemia disease. A Random forest classification is an extension of a decision tree structure, where every inside node specifies a test on a feature, every division characterizes an outcome of the test, and leaf nodes signify classes or class distributions [4]. C4.5 [5] is an algorithm used to produce a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees created by C4.5 can be used for a learning process. C4.5 constructs decision trees from a group of training data in a similar way as ID3, using the perception of information entropy. The support vector machine (SVM) is a newly developed method for multi-dimensional function estimate. The aim of support vector machine is to decide a classifier function which decreases the observed possibility (that is, the training set fault) and the confidence interval (which matches to the overview or test set error) [6] SMO equipments the sequential minimal optimization algorithm for training a support vector classifier using polynomial or Gaussian kernels.

The aim of this work is to develop a capable data mining process of appropriate feature selection and classification is efficient to forecast the iron deficiency between pregnant women to support anemia detection. The paper presents the methodologies namely data preprocessing, improved Median vector feature selection (IMVFS) algorithm and new RandomPrediction (RP) classification algorithm is used to determine the disease type (Non-anemic, Mild and Severe or moderate) prediction is to provide a good forecast outcome and time efficiency. The method does not aim at all kinds of diseases. We just want to contribute only anemia disease for pregnant women and children classification task.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
Dithy M.D, Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India.
Dr. V KrishnaPriya, Professor and Head-PG, School of computing, Sri Ramakrishna College of Arts and Science, Coimbatore, India.

The rest of the paper is organized as follows: Related Work is detailed in Sect. 2. In Sect. 3, research methodologies of data pre-processing feature selection and classification process are to predict the anemia deficiency prediction and experimental results are described in Sect. 4. The conclusion is in Sect. 5.

II. RELATED WORK

***Lee, et.al.,* (2005)** [3] stated that extensive comparison study expanding the performance of newly developed classification techniques in the microarray research, and presented the strategies for searching the majority appropriate classification tools in different situations. They completed their assessment in three ways: additional classification techniques (21 methods), additional datasets (7 datasets) and extra gene prediction methods (3). Their assessment studied showed numerous attractive facts and gives the biologists and biostatisticians some nears into the learning or prediction tools in microarray data examination. The corresponding study also demonstrated that the additional complicated classifiers give enhanced performances than standard techniques such as kNN, DLDA, DQDA and the selection of gene assortment technique has many result on the performance of the learning techniques, and thus the classification techniques must be measured collectively with the gene selection criterion.

Francois, et.al., (2007) [4] discussed a joining the joint information measure with an advanced feature selection policy presents an excellent trade-off among the optimality of the chosen attribute subset and calculation time. But, it needs to group the parameters of shared information estimator and to decide when to stop the forward process. These two options are complicated to make since, as the dimensionality of the separation enlarges, the evaluation of the mutual information turn into fewer and less reliable. The authors proposed to employ re-sampling techniques, K-fold cross-validation, and the variation test, to deal with both problems. The re-sampling techniques carry information about the inconsistency of the estimator, information which can then be used to repeatedly group the parameter and to compute a predefined threshold to end the forward process. The process is demonstrated on a synthetic dataset as well as on real-world instances.

***S. Loscalzo, L. Yu, and C. Ding* (2009)** [5] proposed a new framework for constant feature selection which primary recognizes harmony attribute sets from sub-sampling of training models, and then executes feature or attribute selection by pleasuring each consensus feature set as a distinct entity. Experimentations on together with synthetic and real-world data sets demonstrate that an algorithm created below the framework is successful at improving the issue of undersized instance dimension and guides to additional constant feature selection outcome and comparable or improved simplification presentation than state-of-the-art attribute selection methods.

Iman Azarkhish & Mohammad Reza Raoufy & Shahriar Gharibzadeh (2012) [6] presented an iron deficiency anemia (IDA) is the mainly frequent nutritional deficiency worldwide. Determining serum iron is time-consuming, costly and not accessible in most hospitals. The authors studied, based on four available laboratory data, expanding an artificial neural network and an adaptive neuro-fuzzy inference system to identify the IDA and to forecast, serum iron level. Their outcomes signified that the neural network analysis is greater to ANFIS and logistic regression replicas in identifying IDA. Furthermore, the outcomes demonstrated that the ANN is probable to present a precise test for forecasting serum iron levels with elevated accuracy and suitable precision.

***C. Qiu, L. Jiang, and C. Li,* (2015)** [7] investigated the class possibility assessment performance of SP in terms of CLL and discover that it's class possibility estimation performance approximately ties the novel distribution-based tree augmented naive Bayes. In order to level up its class possibility assessment presentation, authors proposed an enhanced CLL-based SuperParent algorithm. In CLL-SP, a CLL-based advance instead of a categorization based approach, is used to search the expand arcs. The investigational outcomes on a huge group of benchmark datasets showed that the CLL-based approach considerably betters the classification-based approach (SP) and the unique distribution-based approach in terms of CLL, however at the identical time preserves the elevated classification accuracy that distinguishes the classification-based approach (SP).

***Manal Abdullah and Salma Al-Asmari* (2017)** [8] discussed a medical data mining area disturbed with prediction knowledge as a technique to mine preferred results from data for detailed purposes. Anemia is one of the mainly frequent hematological diseases and the study deliberate on the most five frequent kinds of anemia. The authors specified anemia type for the anemic enduring through an analytical model performed some data mining classification methods. The real data of dataset created from the Complete Blood Count (CBC) test outcomes of the patients.

***Dithy M.D and Dr. V KrishnaPriya* (2018)** [1] has discussed to discover out the occurrence of anemia in pregnant women and childish girls and its association of iron deficiency with demographic factors.

The research aimed to present an improved ADD-Left Remove-Right selection (ALRR) sequential feature selection process and Gausnominal classification algorithm to predict the anemia disease lessons (Mild, Not anemic and Severe and moderate) based on the data mining procedures. The outcomes showed that the presentation of the novel method is valuable evaluated with additional Classification of ANN algorithms.

Angshuman Paul, et.al., (2018) [9] proposed an enhanced random forest classifier that executes classification with a least number of trees. Their technique iteratively eliminates various insignificant attributes or features. Based on the amount of significant and insignificant attributes, they created a narrative theoretical higher boundary on the amount of trees to be extra to the forest to guarantee development in classification accuracy. Their algorithm meets with a summary except significant group attributes. The authors proved that additional addition of trees or more reduction of attributes does not progress classification performance. The usefulness of their approach is established through experiments on benchmark data sets. To more use their classifier to notice mitotic nuclei in the histopathological data sets of breast tissues. They also applied their technique on the industrial data set of dual-phase steel microstructures to categorize special phases. Consequences of their techniques on singular data sets showed an important reduction in common classification fault evaluated with an amount of competing methods.

***Dithy M.D and Dr. V KrishnaPriya* (2019)** [2] has discussed an anemia is the common hematological disorder that occurs in pregnancy. The mining of hidden logical information from the huge dataset is probable with data mining process. In presented work, anemia can be recognized in a selection of approaches, based on the Artificial Neural Networks (ANN), Gausnominal Classification, etc…,. In these studies, the appropriate attributes are earlier defined and forecast the classification accuracy. So the authors presented an assessment of the prediction and classification of anemia in patients using data mining methods exclusive of before defined. The work presented an enhanced Roughset based Fuzzy threshold (RFT) feature selection procedure and novel VectNeighbour classification algorithm to predict the anemia disease classes (Mild, Not anemic and Severe and moderate) based on the data mining methods. The outcomes demonstrated that the performance of the novel method is an effective evaluated with additional Classification of ANN and Gausnominal classification algorithms. As the Experimental outcomes showed that VectNeighbour classification with RFT) feature selection techniques obviously better than the existing technique.

III. RESEARCH METHODOLOGY

The proposed research methodology performs the Anemia prediction process using data cleaning or preprocessing, Improved Median Vector Feature Selection and RandomPrediction (RP) Classification process is derived this section. We already introduced a data preprocessing work is presented in [1] and [2], using suitable features with absolute scaling the probability values to convert normalized dataset which is a natural extension of data pre-processing process. The anemia prediction process considers the overall process flow diagram is described in figure 1 (a)&(b).

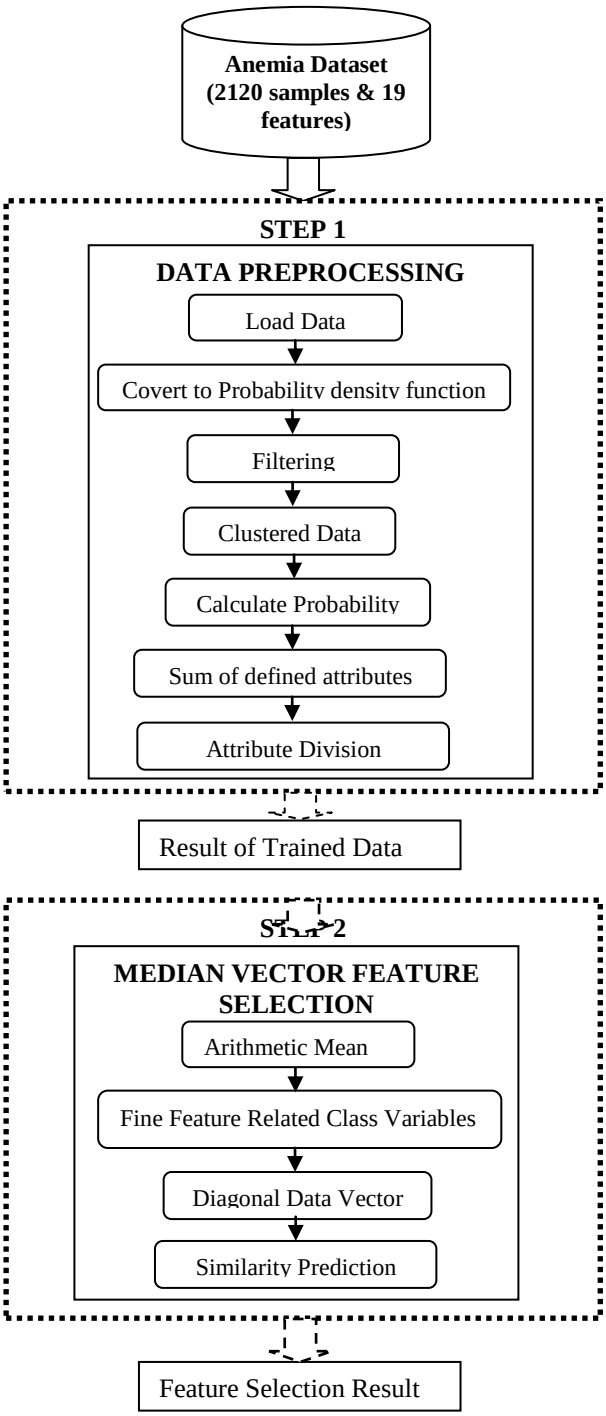


Figure 1 (a). Proposed flow diagram

A. Data Preprocessing

Data cleaning or data preprocessing is an important data mining technique that executes altering original data features into a logical format. The preprocess method try to reduce noise, irrelevant and NaN (Not a Number) that is present in the dataset. In data mining, preprocessing of input data set is knowledge discovery goal approaches usually consume the main portion of the attempt dedicated in the whole work.

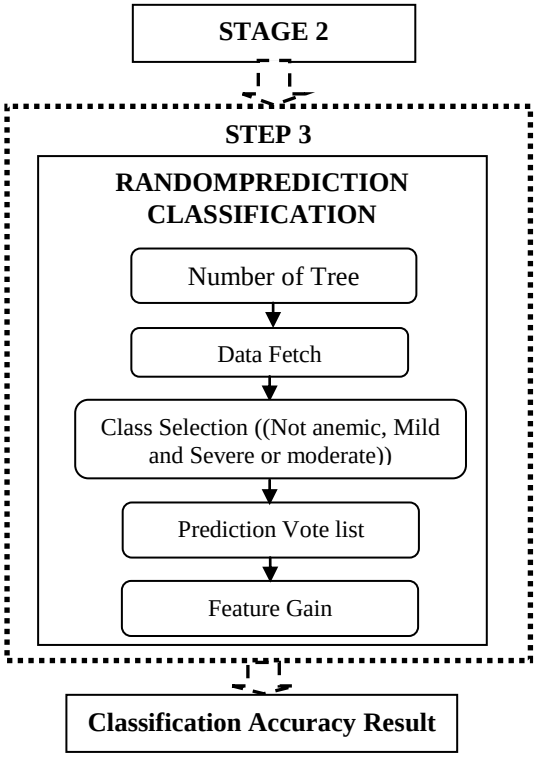


Figure 1 (b). Proposed flow diagram

The present work performs a real-time Anemia dataset has 19 features and 2120 instances contains incomplete, noisy, NaN and inconsistent are described in figure 2.

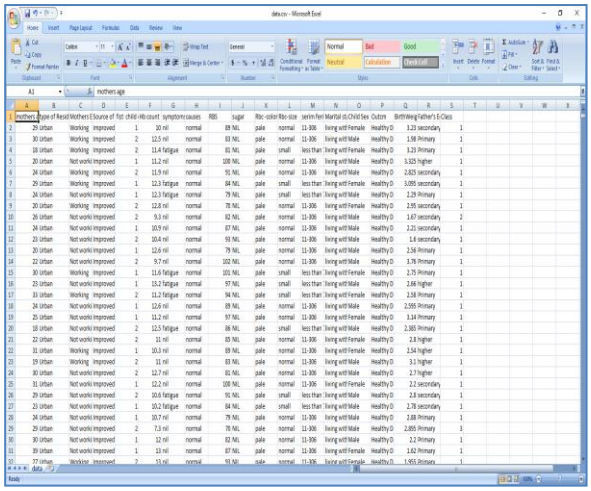


Figure 2. Anemia dataset

After data preprocessing, the entire features are converted in to attribute relation values are described in below figure 3.

B. Improved Median Vector Feature Selection (IMVFS)

An improved median vector feature selection process is derived from the median feature selection technique. The IMVFS process, Forming Median Vector feature selection using the arithmetic mean and data feature similarity from the preprocessed dataset features, optimally according to the defined criterion. This technique is based on the concept of a min and max median vector of diagonal selection class variables.

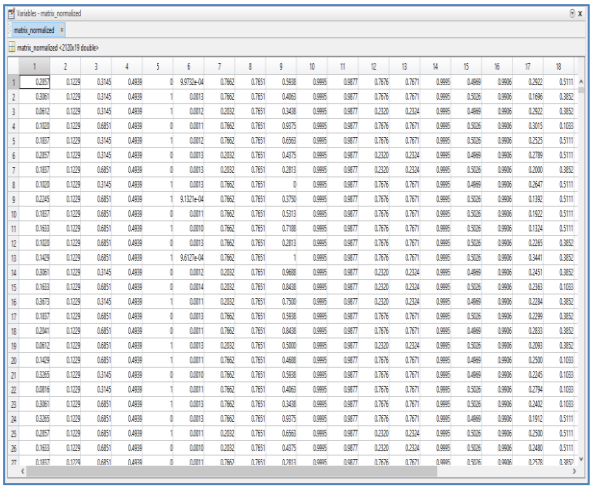


Figure 3. Preprocessed input Anemia dataset

A feature selection system can be represented in Fig. 4,

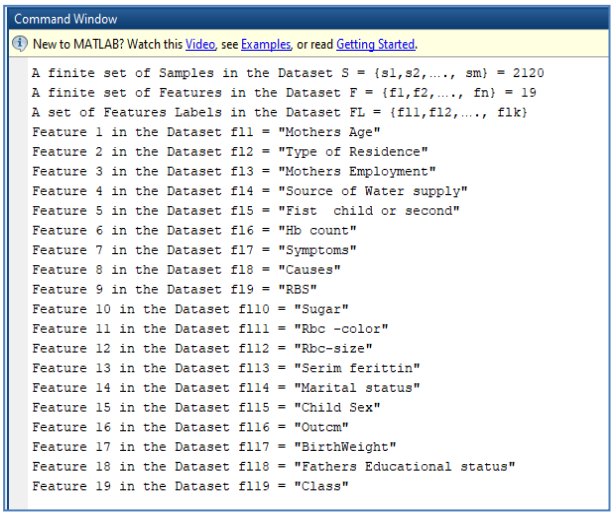


Figure. 4 Number of Features in the Anemia dataset

The IMVFS method can be expressed as,

$$Median(FS) = \sum_{k=1}^{c} \sum_{n=1}^{feat} mean(feat(find(feat = = k), k: feat)) \quad eqn.(1)$$

where c represents class; *feat* is features.

Forming the median vector using the arithmetic mean to discover the relevant feature with appropriate class c.

After discovering the Median(FS), scaling the entire feature with a range of [0, 1] variations along with data feature and class. Equation 2 shows the minimum relevant data vector as,

$$DV = \sum_{n=1}^{f} feat_f + ones * diag(abs(\min(feat_f))) \quad eqn.(2)$$

where *DV* is a set of data feature vector; *f* represents features. To predict the minimum feature selection is not give a complete solution of feature selection. To combine the Median(FS) and DV to get the median vector feature selection as,

MV = *Median* (*FS*)./ max(*DV*) eqn. (3)

The final similarity variations of *MV* is,

$$Sim_{IMFS} = \sum_{j=1}^{m}\sum_{i=1}^{f}\sum_{k=1}^{c} 1 - abs(MV_k^f - feat(j,i)) \quad eqn.(4)$$

Where *m* is number of instances in the dataset; *f* is features; *c* is number of class (*Non-Anemic, Mild and Severe or Moderate*).

Algorithm1: IMVFS Feature selection

Input: Set of features *F*, number of Samples *S*, Class *C*.

Output: A relevant subset of features (*SF*).

Process

Step 1: Initialize SF ← 0;

Step 2: For *j* in 1: number of features (*f*)

Find the Median(FS) using eqn. (1)

End for

Step 3: Forming the data Median Vector (DV) scaling feature range [0, 1] using eqn. (2)

Step 4: To integrate the median(FS) with data vector (DV) as *MV* using eqn. (3).

Step 5: Finally to predict the similarities of *MV* to get the appropriate features (*SF*).

The final IMVFS feature selection results are shown in fig. 5.

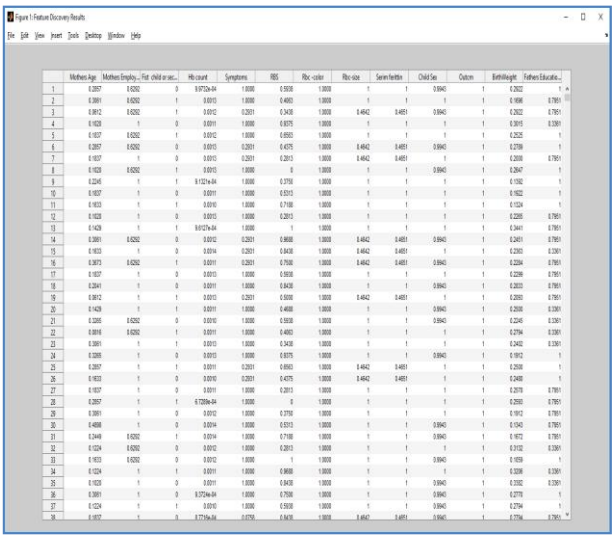


Figure 5. Final IMVFS feature Discovery Result

C. Randomprediction (RP) Classification

The paper presents a novel method of Randomprediction classification is an extension of Random forest and probability classification model. This classification starts with initializing with a number of tree variable and dimension. These classification inputs are final feature selection as test feature; whole preprocessed feature as training feature. This technique preprocesses the training and testing features to obtain the feature sore of feature dimension.

A RM classification case is classified by a number of FOLD (First order logical decision) manner to evaluate the prediction class. The random selection feature classification attains the class (Non-Anemic, Mild and Severe or Moderate) score with their 19 features.

$$Selection_f = \sum_{k=1}^{c}\sum_{j=1}^{f} find(Vote_f = = prediction_c) \quad eqn.(5)$$

Where *f* is features; *c* is number of class (*Non-Anemic, Mild and Severe or Moderate*); Vote is unique features in the trained dataset.

After that selection process, RandomPrediction process is attain the classification accuracy is below equation 6.

$$Perdiction_{acc} = \sum_{m=1}^{len(vote)} Selection(voteIndex_m) \quad eqn.(6)$$

IV. EXPERIMENTAL RESULT

The experimental result has been evaluated the performance of Enhanced Anemia disease prediction with Improved Median Vector Feature Selection (IMVFS) and RandomPrediction (RP) classification. The experimental results execute on Intel I5-6500 series 3.20 GHz 4 core processor, 8GB main memory, and runs on the Windows operating system, on which MATALAB R2013a.

To evaluate the performance of the proposed Classification outperforms the accuracy with existing Gausnominal [1] and VectNeighbour classification [2]. In experimental, the anemia dataset could be applied to proposed RandomPrediction (RP) classification with IMVFS feature selection method which can be flexibly configured to predict the accuracy of data sets to meet the needs of various test requirements.

To evaluate the performance of the proposed Classification model predicts the Anemia classes into three categories (i) Non-Anemic; (ii) Mild; and (iii) Severe or Moderate. Table 1 represents the mother's age wise anemic status.

Table 1. Mothers Age Percentage distribution of Anemia status and Covariates.

Age	Non-Anemic	Mild	Severe or Moderate
18-24	82.1678	10.7517	7.0804
25-29	82.5321	8.3333	9.1346
30-34	81.8533	12.7413	5.4054
35-38	88.6364	9.0909	2.2727

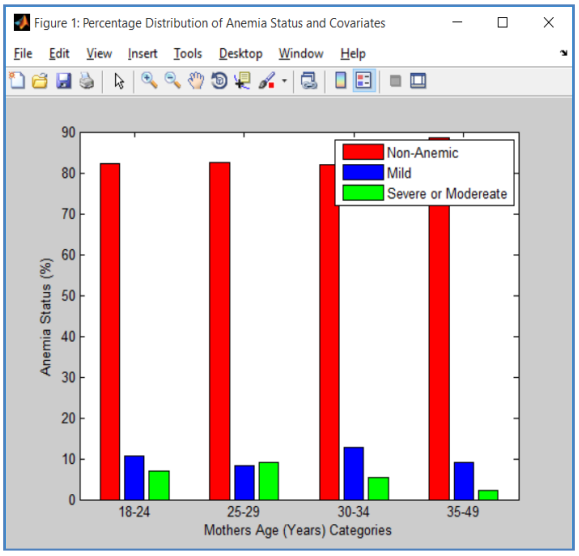


Figure 6. Mothers Age Percentage distribution of Anemia status and Covariates

In figure 7, shows the Mothers age distribution of Anemia status and age wise prediction count is described below,

```
Variables     |  Categories    |  Non-Anemic    |   Mild      |     Severe or Moderate
-------------------------------------------------------------------------------------------
Mothers Age   | 18-24          |    82.1678     |    10.7517  |   7.0804
Mothers Age   | 25-29          |    82.5321     |    8.3333   |  9.1346
Mothers Age   | 30-34          |    81.8533     |    12.7413  |   5.4054
Mothers Age   | 35-49          |    88.6364     |    9.0909   |  2.2727
-------------------------------------------------------------------------------------------
Non Anemic Count Age 18-25 = 940
Mild Count Age 18-25= 123
Severe or Moderate Age 18-25= 81
-------------------------------------------------------------------------------------------
Non Anemic Count Age 25-29 = 515
Mild Count Age 25-29= 52
Severe or Moderate Age 25-29= 57
-------------------------------------------------------------------------------------------
Non Anemic Count Age 30-34 = 212
Mild Count Age 30-34= 33
Severe or Moderate Age 30-34= 14
-------------------------------------------------------------------------------------------
Non Anemic Count Age 35-49 = 78
Mild Count Age 35-49= 8
Severe or Moderate Age 35-49= 2
-------------------------------------------------------------------------------------------
```

Figure 7. Mothers Age Percentage distribution of Anemia status and prediction count

Table 2 represents the type of Residencies (Rural and Urban) distribution of Anemia status and Covariates.

Table 2. Type of Residencies distribution of Anemia status and Covariates.

Type of Residencies	Non-Anemic	Mild	Severe or Moderate
Rural	0	0	0
Urban	82.5	10.1887	7.3113

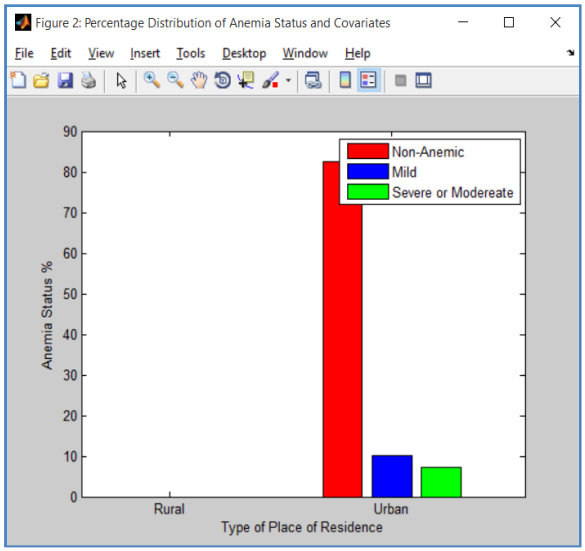


Figure 8. Type of Residencies distribution of Anemia status and Covariates

In figure 8 shows the type of residencies distribution results.

The experiment result, Table 3 compares the Anemia status and Covariates in Mothers age, type of residencies with working and not working category.

Table 3. Comparison the Anemia status and Covariates in Mothers age, type of residencies with working and not working category

Age	Type of Residencies	Mothers Employment	Non-Anemic	Mild	Severe or Moderate
18-24	Urban	Working	84.8739	9.5238	5.6022
25-29	Urban	Working	80.597	9.4527	9.9502
30-34	Urban	Working	85	11.25	3.75
35-38	Urban	Working	88.8889	11.1111	0
18-24	Urban	Not Working	80.9403	11.3088	7.751
25-29	Urban	Not Working	83.4515	7.8014	8.747
30-34	Urban	Not Working	80.4469	13.4078	6.1453
35-38	Urban	Not Working	88.5246	8.1967	3.2787

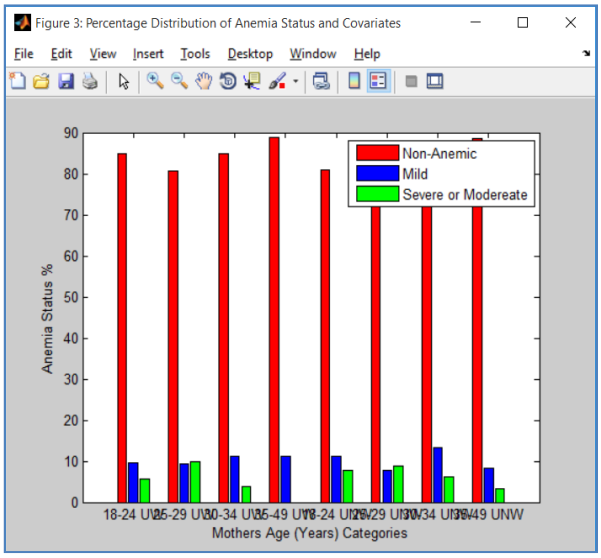
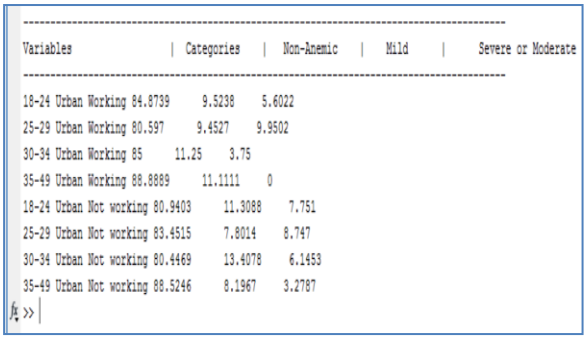


Figure 9. Comparison the Anemia status and Covariates in Mothers age, type of residencies with working and not working category



Variables	Categories	Non-Anemic	Mild	Severe or Moderate
18-24	Urban Working	84.8739	9.5238	5.6022
25-29	Urban Working	80.597	9.4527	9.9502
30-34	Urban Working	85	11.25	3.75
35-49	Urban Working	88.8889	11.1111	0
18-24	Urban Not working	80.9403	11.3088	7.751
25-29	Urban Not working	83.4515	7.8014	8.747
30-34	Urban Not working	80.4469	13.4078	6.1453
35-49	Urban Not working	88.5246	8.1967	3.2787

Figure 10. Anemia status and Covariates in Mothers age, type of residencies with working and not working category

The Proposed RandomPrediction (RP) classification accuracy of 98.94% with compared to our existing work of Gausnominal and VectNeighbour classification algorithms.

$$Accuracy = \frac{Number\ of\ Correctly\ classified\ features}{Number\ of\ instances\ in\ the\ Anemia} \quad (7)$$

Table 4. Overall Classification Accuracy of ANN, Gausnominal, VectNeighbour and Proposed RandomPrediction (RP) classification during randomized run

Methods	1	2	3	4
ANN	0.65	0.69	0.743	0.76
Gausnominal	0.7624	0.852	0.913	0.92
VectNeighbour	0.826	0.896	0.932	0.95
RandomPrediction (RP)	**0.923**	**0.946**	**0.967**	**0.98**

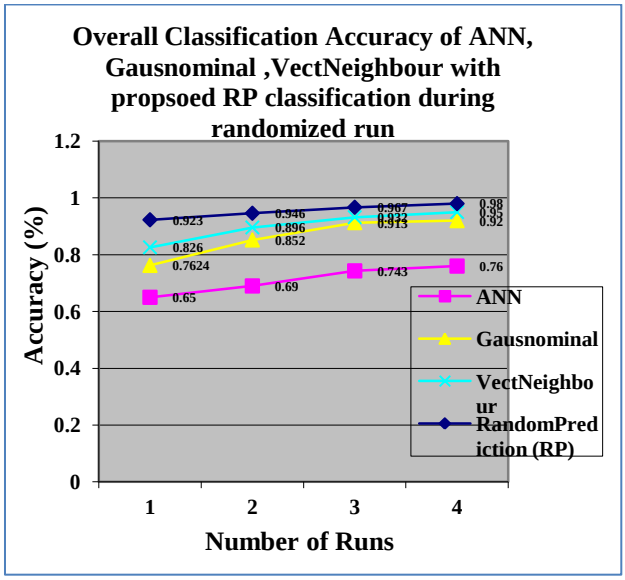


Figure 11. Over all Classification Accuracy Chart

V. CONCLUSION

The paper reviewed and development information of anemia disease discovery in the field of Data preprocessing, IMVFS feature selection and RandomPrediction (RP) classification. The most frequent cause of anemia in pregnancy is iron insufficiency. It usually occurs due to small iron stores previous to pregnancy. This paper design and development with three strategies namely, (1) Data preprocessing; (2) IMVFS feature selection; and (3) RP classification. In the first process, Data preparation is carried out to convert trained statistical data, in the second process, selecting the appropriate feature selection process for predicting the anemia disease and finally predicts the RandomPrediction (RP) classification for predicting accuracy. The Experimental results shows that proposed RandomPrediction (RP) classification with (IMVFS) feature selection methods clearly outperform than previous methods of ANN, Gausnominal and VectNeighbour classification algortithms.

REFERENCES

1. Dithy M.D and Dr. V KrishnaPriya, "Predicting Anemia in Pregnant Women by using Gausnominal Classification Algorithm", International Journal of Pure and Applied Mathematics, Volume 118 No. 20 2018.
2. Dithy M.D and Dr. V KrishnaPriya, "Anemia Screening in Pregnant Women by Using Vect Neighbour Classification Algorithm", Journal of Advanced Research in Dynamical & Control Systems, Vol. 11, 04-Special Issue, 2019.
3. J.W.Lee,J.B.Lee,M.Park,S.H.Song, An extensive comparison of recent classification tools applied to micro array data, Comput.Stat.DataAnal.48(4) (2005) 869–885.
4. D.Francois, F.Rossi, V.Wertz, M.Verleysen, Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70(7–9)(2007)1276–1288.
5. S. Loscalzo, L. Yu, and C. Ding. Consensus group based stable feature selection. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09), 2009, pages 567–576.
6. C. Qiu, L. Jiang, and C. Li, "Not always simple classification: Learning superparent for class probability estimation," Expert Systems with Applications, vol. 42(13), pp. 5433–5440, 2015."

7. Iman Azarkhish & Mohammad Reza Raoufy & Shahriar Gharibzadeh, "Artificial Intelligence Models for Predicting Iron Deficiency Anemia and Iron Serum Level Based on Accessible Laboratory Data", Springer Science+Business Media System, 2012
8. Manal Abdullah and Salma Al-Asmari, "Anemia types prediction based on data mining classification algorithms", Communication, Management and Information Technology – Sampaio de Alencar (Ed.), Taylor & Francis Group, London, 2017
9. Angshuman Paul, Dipti Prasad Mukherjee, Prasun Das, Abhinandan Gangopadhyay, Appa Rao Chintha, Saurabh Kundu, "Improved Random Forest for Classification", IEEE Transactions on Image Processing, Volume: 27, Issue: 8, Aug. 2018.

AUTHORS PROFILE

Ms. Dithy M.D D Pursuing Ph.D. Computer science Full-time Research scholar in Sri Ramakrishna College of Arts and Science for Women at Coimbatore. She did her M.phil Computer science & M.Sc Computer science in kongunadu arts and science college and B.Sc Computer science in V.L.B Janakiammal College of Arts and Science at Coimbatore. she has published 3 papers in international journals and has 3 presentations in International and National conferences.

Dr. V.Krishna Priya is present, professor and head,-PG,school of computing at Sri Ramakrishna College of Arts and Science at Coimbatore. She received her Ph.D. in Computer Science from Mother Theresa University, Kodaikanal. She holds her Masters in Computer Science (MCA) and Bachelors in Chemistry from Bharathiar University. She has published papers in 15 International journals and 15 National Journals and has more than 15 presentations in International and National conferences to her credit. Has 21 years of academic experience and has held various positions at Sri Ramakrishna College of Arts and Science for Women.