Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 154 (2021) 101-116 EISSN 2392-2192

CellProfiler and WEKA Tools: Image Analysis for Fish Erythrocytes Shape and Machine Learning Model Algorithm Accuracy Prediction of Dataset

Soumendra Nath Talapatra[1,*], **Rupa Chaudhuri**[2] and **Subhasis Ghosh**[2]

[1]Department of Bio-Science, Seacom Skills University,
Kendradangal, Birbhum, West Bengal, India

[2]Department of Environmental Science, University of Calcutta,
35 Ballygunge Circular Road, Kolkata, India

*E-mail address: soumendrat@gmail.com

ABSTRACT

The first part of the study was detected the number of cells and measurement of shape of cells, cytoplasm, and nuclei in an image of Giemsa-stained of fish peripheral erythrocytes by using CellProfiler (CP, version 2.1.0) tool, an image analysis tool. In the second part, it was evaluated machine learning (ML) algorithm models viz. BayesNet (BN), NaiveBayes (NB), logistic regression (LR), Lazy.KStar (K*), decision tree (DT) J48, Random forest (RF) and Random tree (RT) in the WEKA tool (version 3.8.5) for the prediction of the accuracy of the dataset generated from an image. The CP predicts the numbers and individual cellular area shape (arbitrary unit) of cells, cytoplasm, and nuclei as primary, secondary, and tertiary object data in an image. The performance of model accuracy of studied ML algorithm classifications as per correctly and incorrectly classified instances, the highest values were observed in RF and RT followed by K*, LR, BN and DTJ48 and lowest in NB as per training and testing set of correctly classified instances. In case of performance accuracy of class for K value, the highest values were observed in RF and RT followed by K*, LR, BN and DTJ48 and lowest in NB while lowest values were obtained for mean absolute error (MAE) and root mean squared error (RMSE) in case of RT followed by RF, K*, LR, BN and DTJ48 and comparatively highest value in case of NB as per training and testing set. In conclusion, both tools performed well as an image to the dataset and obtained dataset to rich information through ML modelling and future study in WEKA tool can easily be analysed many biological big data to predict classifier accuracy.

(Received 22 January 2021; Accepted 12 February 2021; Date of Publication 13 February 2021)

***Keywords*:** Fish erythrocytes quantification, Shapes measurement, Automatic image analysis, CellProfiler tool, Machine learning algorithm, Model classifier accuracy, WEKA tool

1. INTRODUCTION

The architecture of any cell can easily be determined under the microscope after using appropriate stain. This differential staining help to know the normal or abnormal shape of cells and nuclei. The measurement of the cells after calculating through ocular micrometre and calibrated by stage micrometre. A technique for the shape measurement of the erythrocyte was developed by Adams (1954) followed by the study of red blood cells (Diez-Silva et al., 2010).

Moreover, the study of fish blood cells morphometry is determined the pathological features (Baak, 1985; Vandiest and Baak, 1991; Acharya and Mohanty, 2014; Pala and Dey, 2016; Shen et al., 2018).

But this measurement is time-consuming, individual eye estimation error, missing of cellular or nuclear features and also required proper expertise to clarify cellular and nuclear features under the microscope. Manually, the measurement of the size of the cells and nuclei is very tedious work and possibilities of error in data analysis. In recent days, measurement of the shape of cells and nuclei by using high throughput tools are showing less time, no visual error, proper measurement of cells and nuclei, etc. The cell measurement through algorithm-based software has been recommended by many researchers (Wahlby et al., 2004; Carpenter et al., 2006; Lamprecht et al., 2007; Jones et al., 2009; Ljosa and Carpenter, 2009; Kamentsky et al., 2011; Bray et al., 2015; Talapatra et al., 2016) but the task is careful setting up of input in the tool (Bray et al., 2015).

Generally, image analysis to know accurate identification and measurement of all types of cellular features through automatic analysis of certain phenotypes through software, previously studied by many researchers (Carpenter et al., 2006; Jones et al., 2009; Bray et al., 2015; Talapatra et al., 2016). According to Carpenter et al. (2006), an image analysis software, CP is a freely available image analysis tool, can be capable of handling 100 nos. of images of any cell types like yeast colony to mammalian cells (Bray et al., 2015). CP software helps to detect the biological information quickly with statistical power and the simple cellular morphology viz. cell count and shape in an individual cell as well as complex morphological parameters like cell/organelle shape or subcellular patterns of DNA or protein from stained images. Many works have been carried out on fluorescently stained cells, but earlier study attempted before Giemsa-stained image analysis from few normal and abnormal erythrocytes of fish to detect rich information of data from cells by using CP (Talapatra et al., 2016).

Interestingly, in recent decades, big data mining is challenging research in which the endeavour from dataset to rich information. This can easily be achieved through machine learning (ML) modelling or artificial intelligence (AI) algorithms, which predict the performance accuracy of the data (LeCun et al., 2015; Mishra et al., 2021). Besides several big data analysis on finance, agriculture, biomedical science, etc. (Bhuvaneswari and Sarma Dhulipala, 2013; Merelli et al., 2014; Hamid and Ahmed, 2016; Bhatia et al. 2017; Chakraborty et al., 2017; Almryad and Kutucu, 2020; Attwal and Dhiman, 2020; Mishra et al., 2020; Mishra et al., 2021), the data analysis of biological origin is the recent research interest, and recently many biologists are showing interest on the big data analysis by using ML and AI classification algorithms to obtain the accuracy in the big dataset (Chakraborty et al., 2017; Almryad and Kutucu, 2020).

In the first part of the study was detected the number of cells and measurement of cells, cytoplasm, and nuclei in an image of Giemsa-stained peripheral erythrocyte of fish by using CP (version 2.1.0), an image analysis tool. In the second part of the study, it was evaluated machine learning (ML) classification models in the WEKA (Waikato Environment for Knowledge Analysis) tool (version 3.8.5) for the prediction of the accuracy of the dataset generated from an image.

2. MATERIALS AND METHODS

2. 1. Study of image processing and gathering of dataset from an image

An image of Giemsa-stained peripheral erythrocytes of fish was processed by using CP (Version 2.1.0) tool. This tool was retrieved from the designated website (http://www.cellprofiler.org/download.shtml). The input data were incorporated in the present tool with the help of published CP manual as per the link (http://www.cellprofiler.org/linked_files/Documentation/cp2.1.1_manual_6c2d896.pdf) for detail description for users (Carpenter et al., 2006) and earlier study (Talapatra et al., 2016).

The CP software input data was selected from earlier work by Talapatra et al. (2016). In the present study, an image was selected to detect a total number of cells along with cytoplasm and nuclei, and also measured the area (arbitrary unit) to know normal and abnormal erythrocytes of fish as per the earlier study of Talapatra et al. (2016). All the data were obtained through an image as per several computational simulations and saved as .csv file.

2. 2. Study of data mining and predicting information of dataset by machine learning modelling algorithm

In the present study, data mining through ML modelling algorithm was performed by using WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5) developed by Frank et al. (2016). The WEKA explorer was developed with data pre-processing, classification, regression, and association rules. In pre-processing, all the data were made through unsupervised instance followed by segregation of the data as the training set (60%) and rest data (40%) was used as cross-validation (CV) data and 50% of these CV data was used as test set (Bhuvaneswari and Sarma Dhulipala, 2013).

The predictive accuracy of data of fish erythrocytes on the shape of cells, cytoplasm, and nuclei through ML modelling algorithms especially different classifiers viz. BayesNet (BN), NaiveBayes (NB), logistic regression (LR), Lazy.KStar (K*), decision tree (DT) J48, Random forest (RF) and Random tree (RT) along with 4 attributes viz. cells, cytoplasm, nuclei, effect (normal and abnormal structure) from dataset to know the overall performance accuracy.

The performance of model accuracy of above-mentioned ML algorithm classifications as per correctly and incorrectly classified instances, Kappa (K) statistics, mean absolute error (MAE) and root mean squared error (RMSE) were studied for training set and testing set. As per Bouckaert et al. (2020), the modelling summary of results was considered and retrieved from WEKA tool. The prediction accuracy of studied ML models as per training and test set was retrieved from summary results and the statistical parameters are true positive (TP), false positive (FP), Matthews correlation coefficient (MCC), receiver operating characteristic (ROC) and Precision-recall curve (PRC), respectively.

3. RESULTS AND DISCUSSION

The image of Giemsa-stained peripheral erythrocytes of fish was used as an input image along with setting up of selected parameters such as correct illumination application and calculation, cropping, edge enhancement, masking of an image, primary objects (nuclei) identification, nuclei masking and filtering in the image, secondary objects (cells) and tertiary objects (cytoplasm) identification in the image. The original image was obtained through the CP in measured position after incorporated in the CP software. All these parameters were used as input criteria in the CP software as described in the previous study (Talapatra et al., 2016). In the image, a total of 232 nos. of objects were identified by CP tool and the data of area shape (arbitrary unit) were obtained separately for cells (Figure 1), cytoplasm (Figure 2), and nuclei (Figure 3).

The present study was screened easily objects in the image of peripheral erythrocytes of fish after automated analysis through CP, an image analysis software. This software can be an alternative of scoring tool for microscopic images in which the quantification of cells as well as the measurement of the shape of the cell structure through high throughput way (Carpenter et al., 2006). Moreover, the shape of the cells can measure manually under a microscope but the measurement of the cells, cytoplasm, and nucleus separately a tedious job and may have possibilities of errors. According to the concept and software manual, the present study was evaluated as an image containing numerous objects. The cellular and nuclear abnormalities data obtained after the analysis of CP software, which is a faster screening tool to know the shape of the cells, cytoplasm, and nuclei in each cell type in numerous cells population. Internationally, CP has already been approved by several laboratories for dry lab works.

The researchers are studying a variety of biological processes in different cell types of the rat model, fish erythrocytes, etc. (Carpenter et al., 2006; Bray et al., 2015; Talapatra et al., 2016). The present study with Giemsa-stained erythrocytes was screened very easily through quantification and shape measurement properly by using this software. It is required necessary suggestions by other researchers, those who have been investigated and/or working with fluorescence-stained cells.

All these data were retrieved from the software (CP) and used for the building of ML models to predict the accuracy of these big dataset in WEKA tool. In the pre-processing step, graphical representation of statistical data of different attributes (cells, cytoplasm, and nuclei as well as effect) was obtained (Figure 4A-D). Moreover, visual qualitative and quantitative understanding of the distribution class (class effect normal as blue coloured and abnormal as red coloured cells nominal) is depicted in Figure 4A-D in which cells attribute were found ranged between 1.65-5.41 in X axis and the frequency values (%) were obtained 5-57 in Y axis (Figure 4A), cytoplasm attribute was found ranged between 1.00-5.17 in X axis and the frequency values (%) were obtained 0-217 in Y axis (Figure 4B), nuclei attribute was found ranged between 1.48-5.25 in X axis and the frequency values (%) were obtained 4-51 in Y axis and effect attribute viz. normal and abnormal cells were obtained 86 nos. and 146 nos. respectively.

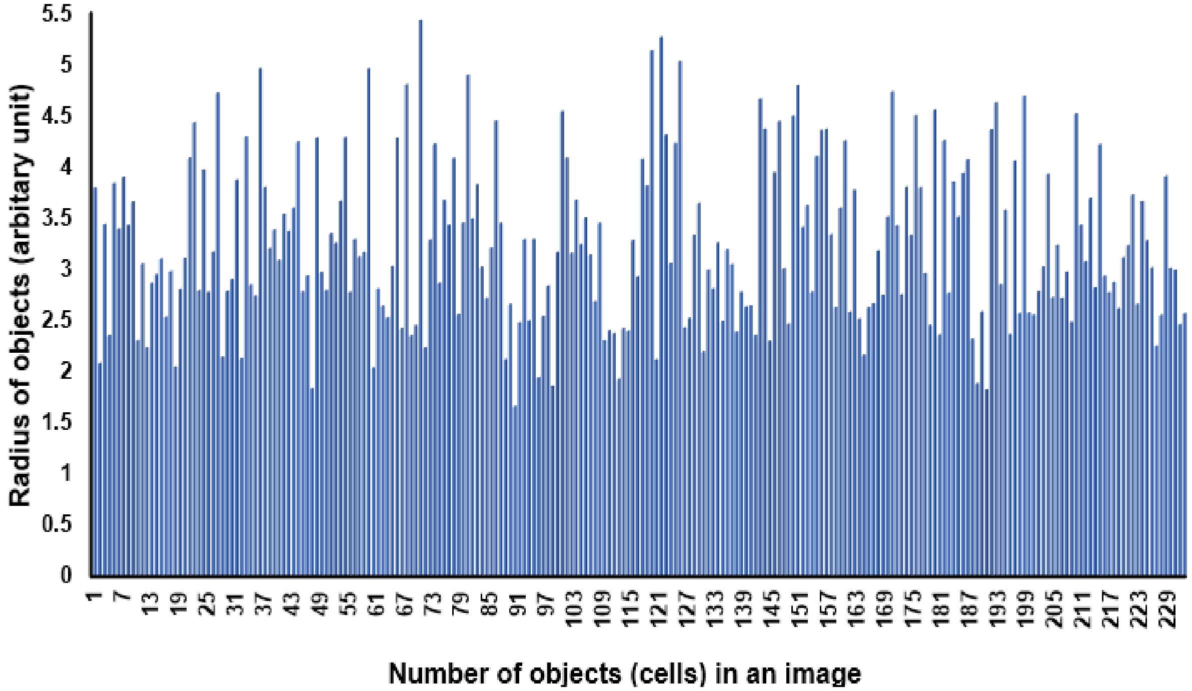


Figure 1. Histogram of area shape of objects especially cells (arbitrary unit) in an image

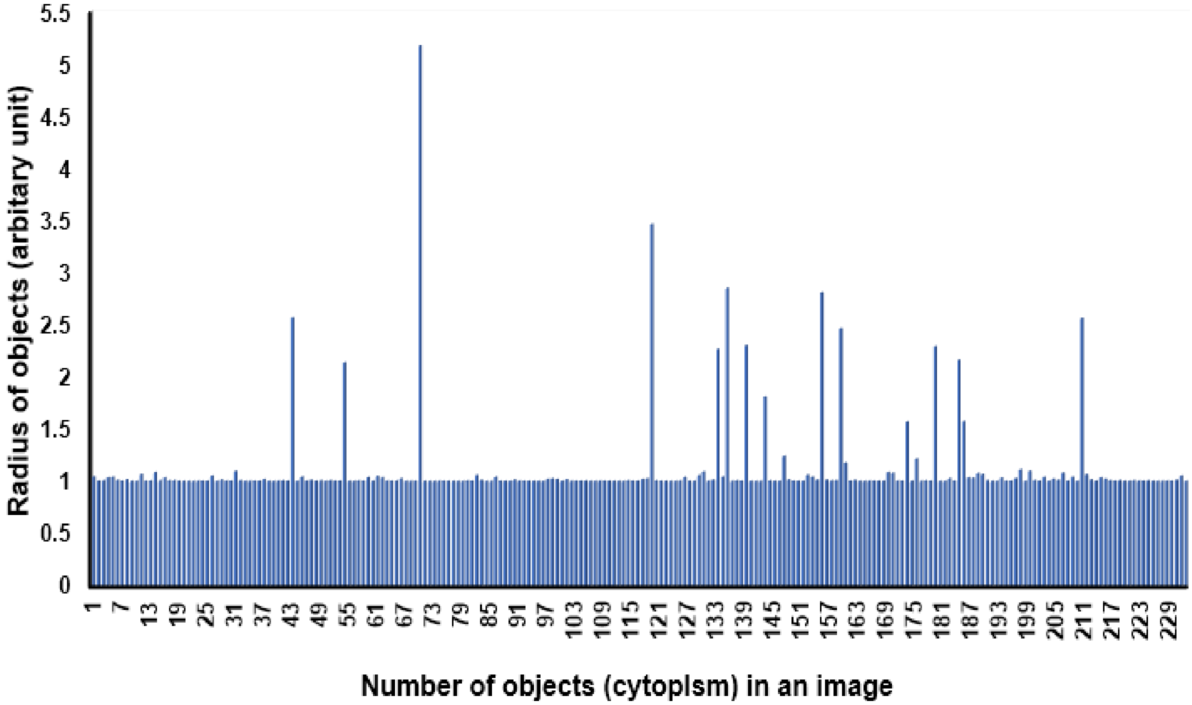


Figure 2. Histogram of area shape of objects especially cytoplasm (arbitrary unit) in an image

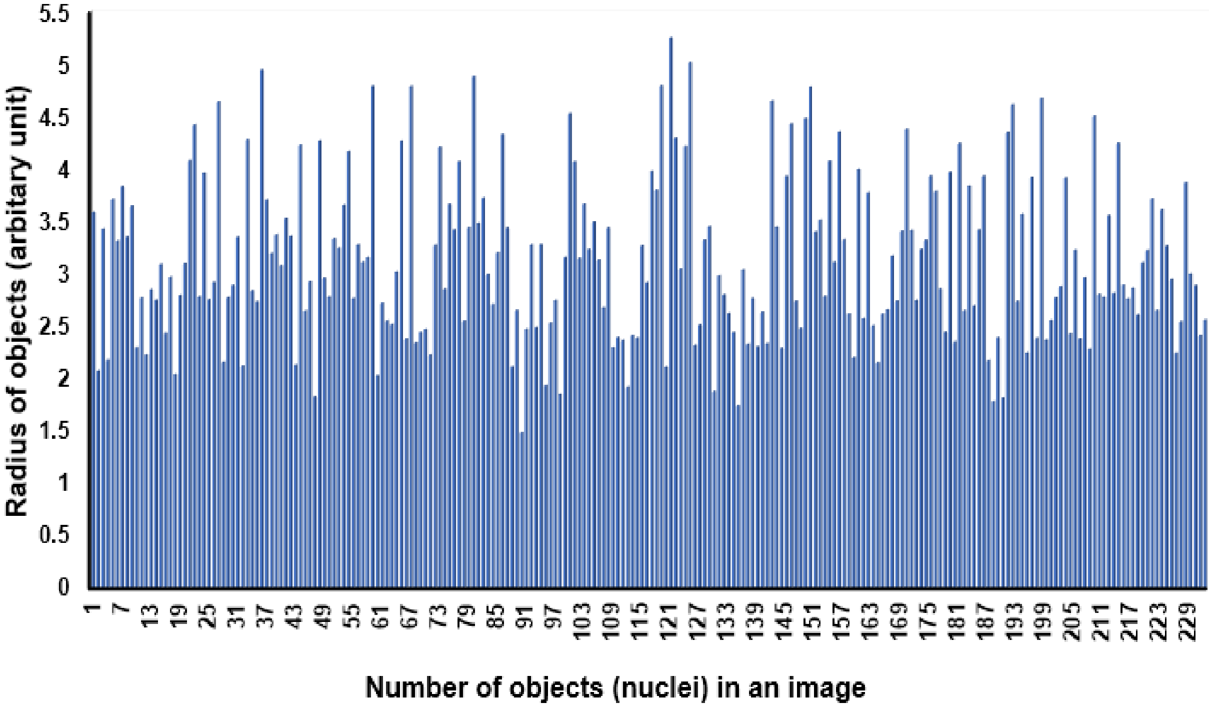
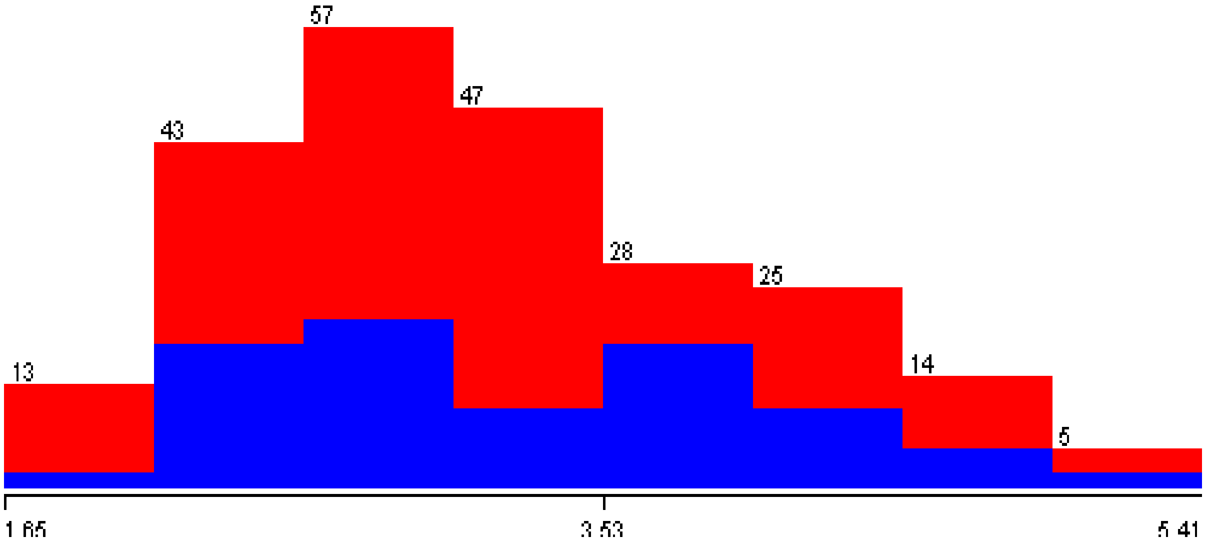
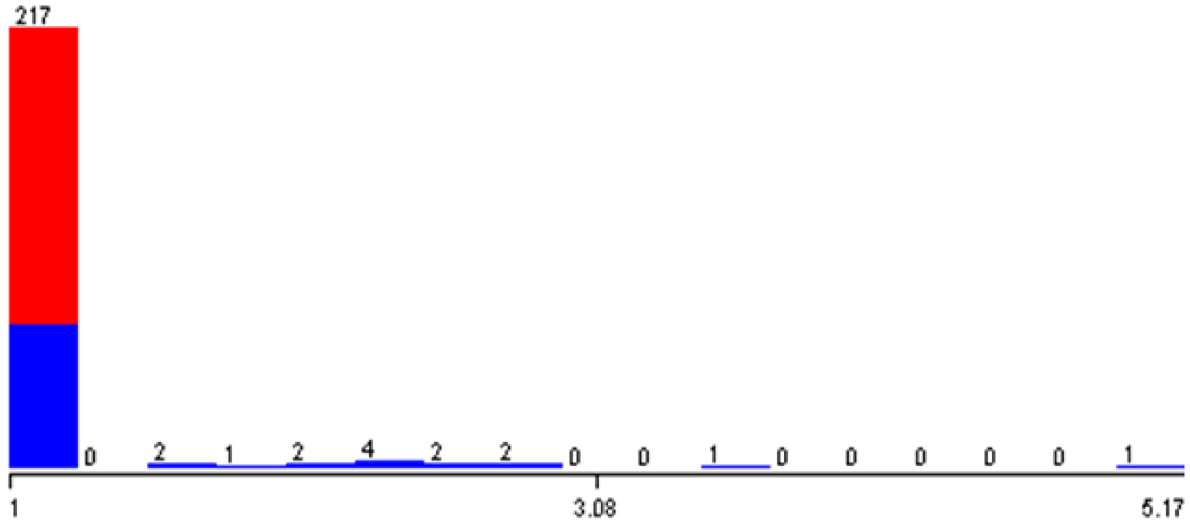


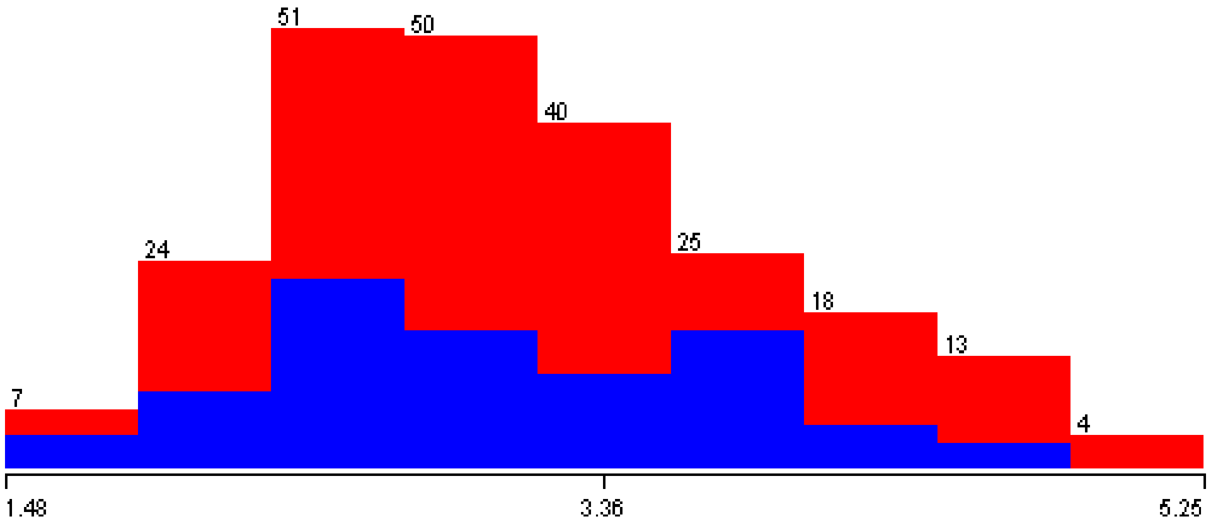
Figure 3. Histogram of area shape of objects especially nuclei (arbitrary unit) in an image



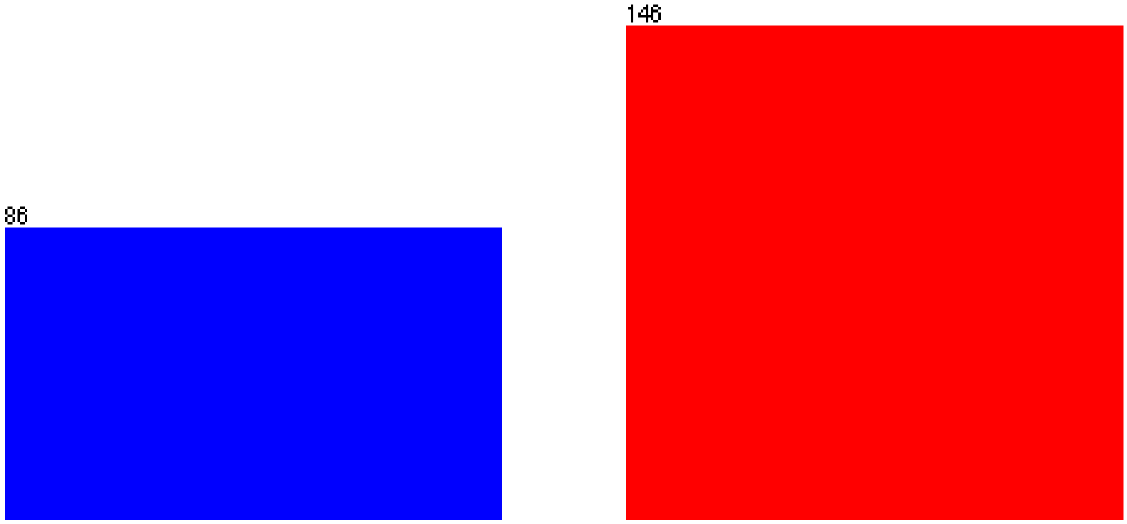
A. Cells attribute



B. Cytoplasm attribute



C. Nuclei attribute



D. Effect attribute

Figure 4 (A-D). Representation of different attributes after pre-processing in WEKA tool

Table 1 describes the statistical values of three different attributes. In case of cells, cytoplasm and nuclei, the mean ± standard deviation (SD) values were obtained 3.32 ± 0.79, 1.11 ± 0.43 and 3.12 ± 0.77, respectively. The weight of effect attribute viz. normal and abnormal cells was obtained 86 and 146, respectively.

Table 1. Statistical values of different attributes

Statistics	**Cells**	**Cytoplasm**	**Nuclei**
Minimum	1.65	1.00	1.45
Maximum	5.41	5.17	5.25
Mean	3.32	1.11	3.12
SD	0.79	0.43	0.77

SD = Standard deviation

Table 2 describes the summary results of correctly and incorrectly classified instances of studied models related to the training set and testing set. In the case of algorithm classification, the highest values were observed in RF and RT followed by K*, LR, BN and DTJ48 and lowest in NB as per training and testing set.

Table 3 describes the summary results of Kappa (K) statistic, mean absolute error (MAE) and root mean squared error (RMSE) of studied models related to the training set and testing set. In case of performance accuracy of the class of K value, the highest values were observed in RF and RT followed by K*, LR, BN and DTJ48 and lowest in NB while lowest values were obtained for MAE and RMSE in case of RT followed by RF, K*, LR, BN and DTJ48 and comparatively highest value in case of NB from other algorithms as per training and testing dataset.

Table 2. Summary results of different models (correctly and incorrectly classified instances)

Classifier model	Correctly classified instances		Incorrectly classified instances	
	TrS (%)	TS (%)	TrS (%)	TS (%)
BN	92.80	93.48	7.19	6.52
NB	79.14	78.26	20.86	21.74
LR	94.96	95.65	5.04	4.35
K*	96.40	95.65	3.60	4.35
DT (J48)	92.80	93.48	7.19	6.52
RF	100.0	100.0	0.00	0.00
RT	100.0	100.0	0.00	0.00

BN = BayesNet; NB = NaiveBayes; LR= Logistic regression; K* = Lazy.KStar; DT (J48) = Decision tree; RF = Random forest; RT = Random tree; TrS = Training set; TS = testing set

Table 3. Model summary (Kappa statistic, mean absolute error and root mean squared error) results.

Classifier model	TrS	TS	TrS	TS	TrS	TS
	KS		MAE		RMSE	
BN	0.85	0.86	0.14	0.13	0.26	0.25
NB	0.53	0.49	0.22	0.23	0.45	0.46
LR	0.89	0.91	0.08	0.07	0.20	0.20
K*	0.92	0.91	0.09	0.09	0.16	0.16
DT (J48)	0.85	0.86	0.13	0.13	0.26	0.25
RT	1.00	1.00	0.00	0.00	0.00	0.00
RF	1.00	1.00	0.05	0.06	0.11	0.11

BN = BayesNet; NB = NaiveBayes; LR= Linear regression; K* = Lazy.KStar; DT (J48) = Decision tree; RF = Random forest; RT = Random tree; KS = Kappa statistic; MAE = Mean absolute error; RMSE = Root mean squared error; TrS = Training set; TS = test set

Figure 5 exhibits a random tree, which is formed by a stochastic or random process. In this case, each node is cleaved through the best split among all attributes and each node (attribute) is further cleaved into the subset of randomly chosen attribute at that node. It is indicated that three trees cleaved from the chosen parameters and reduced the size at the end of the tree. In the present observation 17 nodes and 18 leaves along with branches. The first splitting nodes were observed for cytoplasm at <1.01 and ≥1.01, the second splitting was obtained for cytoplasm at <1.02 and ≥1.02 were found connected with leaves and further ended with branches right side at <1.00 and ≥1.00 while the third splitting was noted for nuclei at <2.34 and ≥2.34 and fourth splitting was found for cell at <2.31 and ≥2.31 ended with leaves and nuclei at <3.85 and ≥3.85 further cleaved into leaves and nodes of cell at <3.91 and ≥3.91 ended with leaves and further cleaved into cell at <4.79 and ≥4.79 ended with leaves. There was found centrally located splitting at <1.00 and ≥1.00 ended right side with leaves and further nuclei cleaved at <2.97 and ≥2.97 ended left side with leaves and further cell cleaved at <3.20 and ≥3.20 ended with leaves left side and further cytoplasm cleaved at <1.00 and ≥1.00 ended with leaves left side and finally, cytoplasm ended into leaves at <1.00 and ≥1.00.

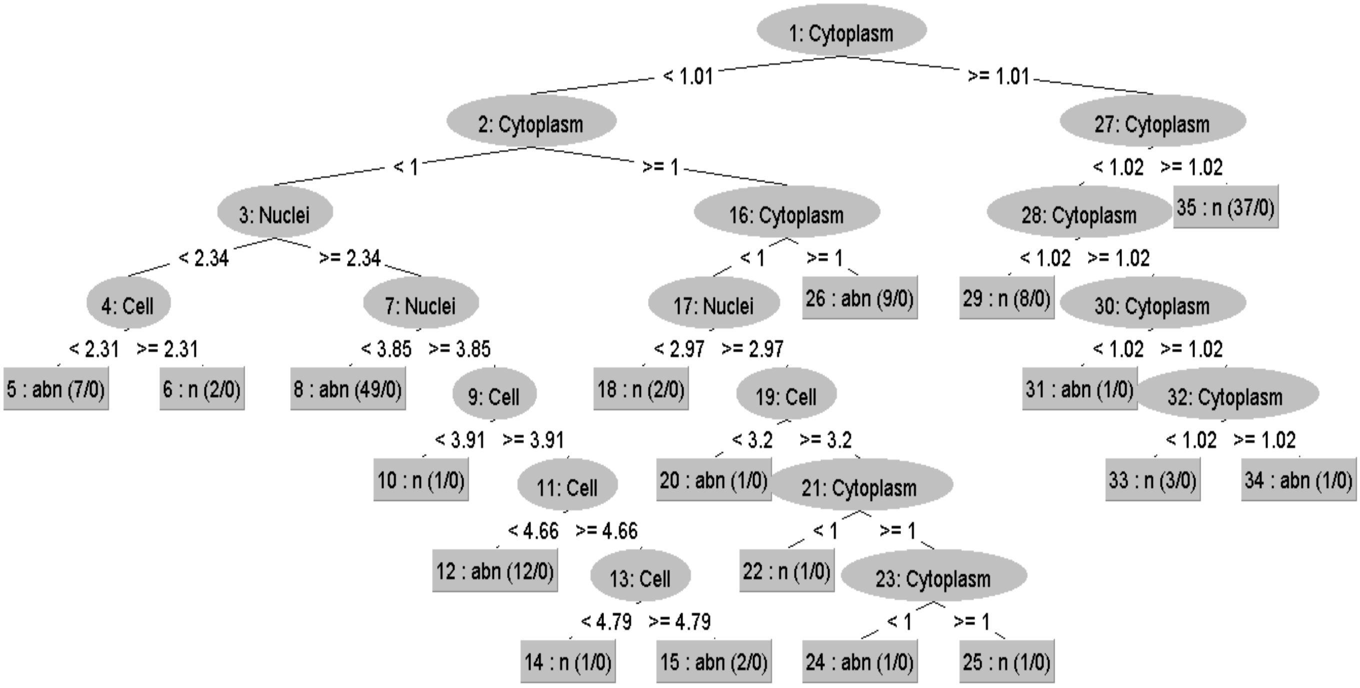
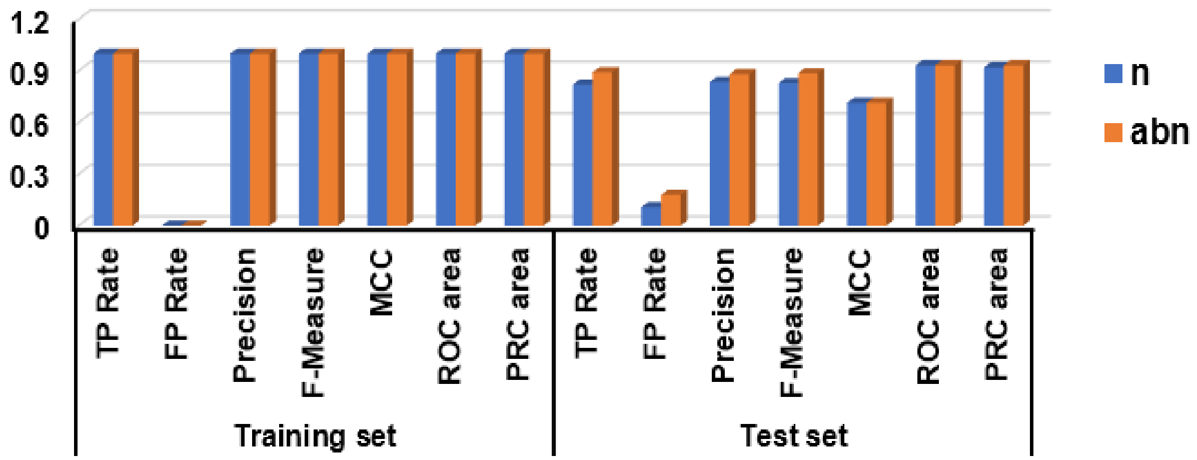
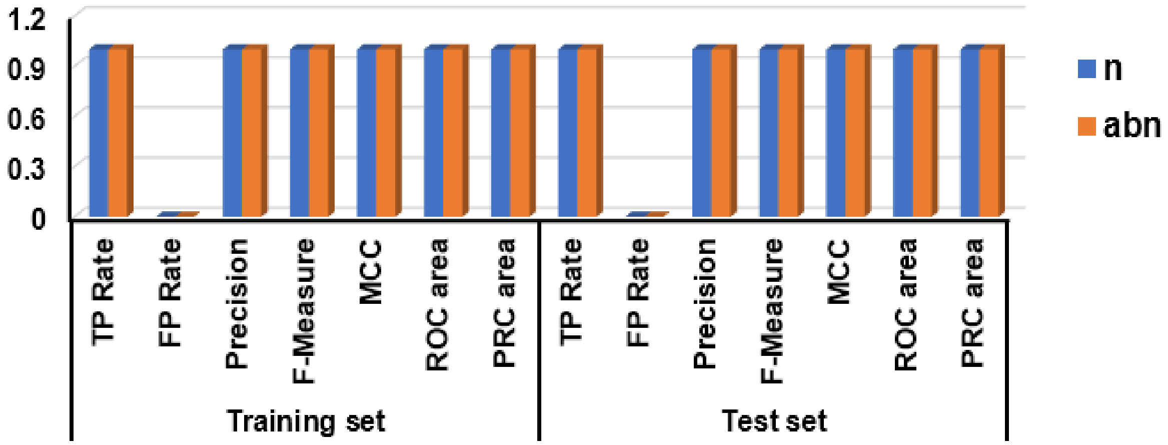


Figure 5. Representation of random tree as per training and test set

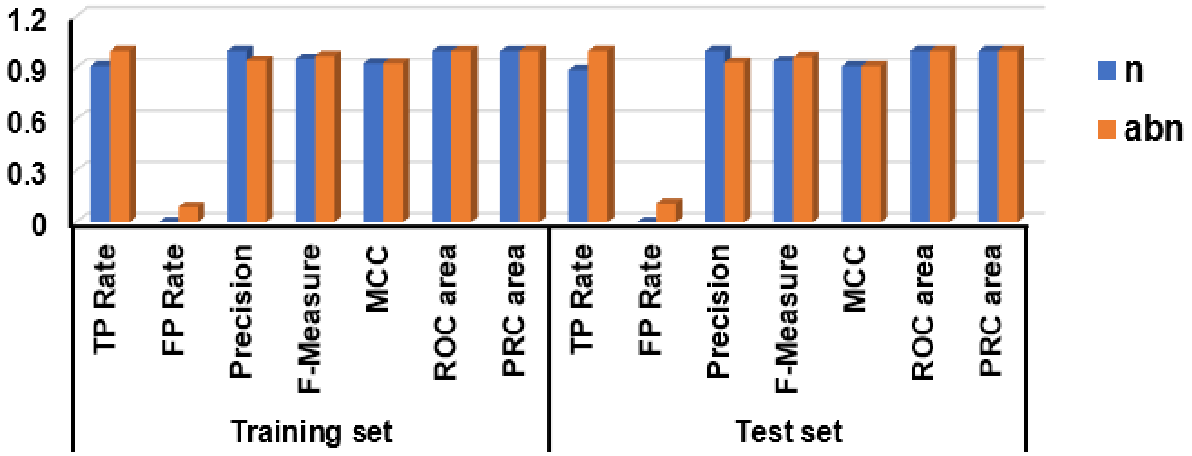
Figure 6 describes the graphical representation of the detailed accuracy of studied models related to the training set and test set. In case of the accuracy of a class of values of TP, FP, precision, MCC, ROC and PRC, the highest values were observed in RF and RT followed by K*, LR, BN and DTJ48 and lowest in NB algorithms as per training and testing dataset. To determine the correctness of the classification (Figure 6A-G), the rate of TP and FP are important statistical parameters of a test. In the present study, a higher rate of TP and lower rate FP was observed in the classification algorithms viz. RF and RT followed by K*, LR, BN and DTJ48 in training and testing dataset. Herein, the ROC curve is observed closely related to TP and FP rate and higher value of the ROC curve is also determined in above-mentioned models except for NB classifier. The MCC is also an important statistical parameter to determine the good score in the prediction result and higher MCC values are observed in above-mentioned models except for NB classifier. In binary classification modelling, higher PRC value is also determined the performance accuracy in the dataset. The present study found better PRC (>95%) values in RF and RT followed by K*, LR, BN and DTJ48 in training and testing dataset, which is accepted in ML algorithm models.



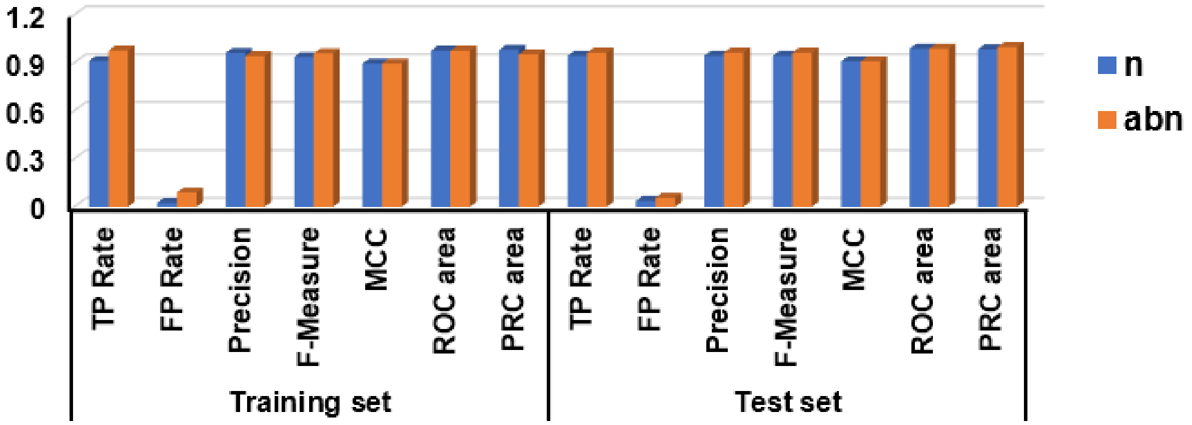
A. Statistical data for performance accuracy of Random Forest algorithm



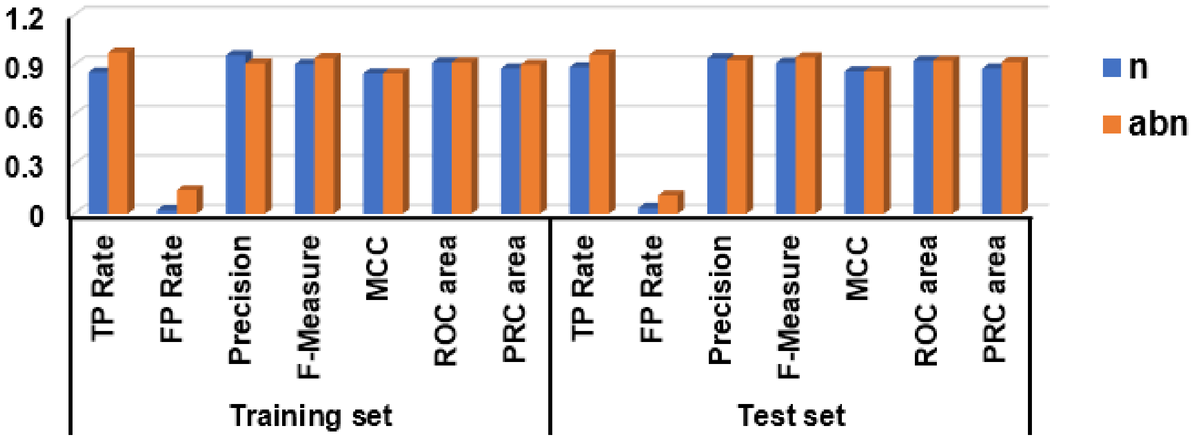
B. Statistical data for performance accuracy of Random Tree algorithm



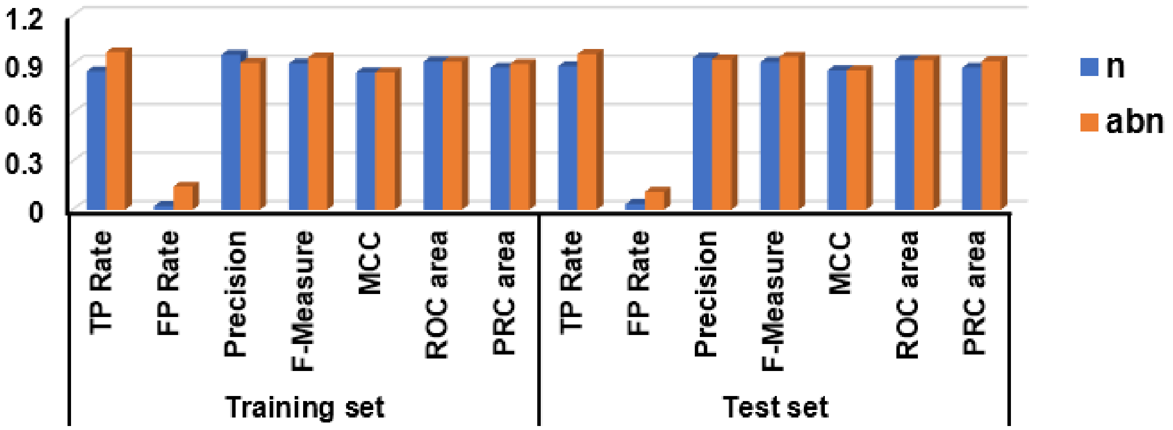
C. Statistical data for performance accuracy of K* algorithm



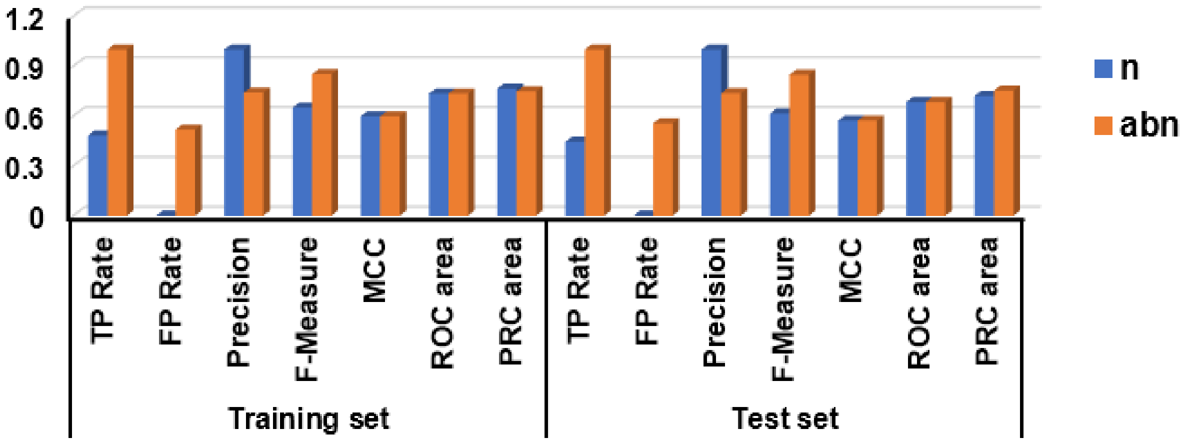
D. Statistical data of for performance accuracy Multinomial Logistic Regression algorithm



E. Statistical data for performance accuracy of Bayes Network Classifier



F. Statistical data for performance accuracy of decision Tree J48 Classifier



G. Statistical data for performance accuracy of Naïve Bayes Classifier

Figure 6 (A-G). Graphical representation of prediction accuracy of different ML models as per training and test set (TP = True positive; FP = False positive; MCC = Matthews correlation coefficient; ROC = Receiver operating characteristic; PRC = Precision-recall curve)

The WEKA tool determines ML classification algorithm models, which helps in data mining of big dataset. The performance accuracies of training and testing dataset related to classified instances, error rate, and kappa statistics and precision statistics values can be predicted easily within this tool (Bhuvaneswari and Sarma Dhulipala, 2013; Merelli et al., 2014; LeCun et al., 2015; Hamid and Ahmed, 2016; Bhatia et al. 2017; Chakraborty et al., 2017; Zainudin et al., 2019; Almryad and Kutucu, 2020; Attwal and Dhiman, 2020; Mishra et al., 2020). In an earlier study of biological and biomedical data mining was used in WEKA tool. Bhuvaneswari and Sarma Dhulipala (2013) studied different ML classifier models such as

Naïve Bayes (NB), Support Vector Machine (SVM), K-nearest neighbours (IBK) and Decision tree (DT) J48 for animal kingdom dataset, Jamil (2016) was used NB, sequential minimum optimization (SMO), DTJ48, K* and Artificial Neural Networks (ANNs) for breast cancer dataset, Zainudin et al. (2019) was studied on convolutional neural network (CNN) algorithm. Siraj-Ud-Doulah (2019) studied on bioinformatics through several ML and AI modelling in which common classifiers viz. DT, J48, K*, LR, NB, RF are found related to the present study.

Some recent studies on ML and AI algorithms have been carried out on finance (Hamid and Ahmed, 2016; Bhatia et al. 2017), biological science (Bhuvaneswari and Sarma Dhulipala, 2013; Zainudin et al., 2019; Attwal and Dhiman, 2020; Almryad and Kutucu, 2020), bioinformatics (Merelli et al., 2014; Siraj-Ud-Doulah, 2019), biomedical science (Asaoka et al., 2017; An et al., 2019; Mishra et al., 2020; Mishra et al., 2021), agriculture (Meeradevi et al., 2020), etc. but combined study with two tools related to generation of the dataset from an image in CP tool and analysis of dataset through ML modelling algorithm to predict the classifier performance accuracy in WEKA tool is the first-time endeavour.

4. CONCLUSIONS

It is concluded that present study is an approach to screen image of Giemsa-stained peripheral erythrocytes of fish to know the number of cells and shape of the cells based on object identification and its shape especially cell, cytoplasm, and nucleus by using CP tool, an automated image-based analysis software. This study helped the computational biological research for extraction of rich information of dataset in the image and to easily know the shape especially size of cells, cytoplasm, and nuclei. On the other hand, the rich information from studied dataset through ML modelling classifiers are obtained better performance accuracy of algorithms viz. RF and RT followed by K*, LR, BN and DTJ48 and lowest in NB as per training and testing dataset. It is suggested in a future study with WEKA tool for other biological big data to predict classifier performance accuracy.

Acknowledgement

Authors convey thanks to all the developers of the present tools used in this study.

References

[1] Adams GC, A technique for the measurement of erythrocyte diameters. *Journal of Clinical Pathology* 7 (1954) 76

[2] Diez-Silva M, Dao M, Han J, Lim CT, Suresh S, Shape and biomechanical characteristics of human red blood cells in health and disease. *MRS Bulletin* 35(5) (2010) 382-388

[3] Baak JPA, The principles and advances of quantitative pathology. *Analytical and Quantitative Cytology and Histology* 9 (1985) 89-95

[4] Vandiest PJ, Baak JPA, Morphometry, In: Comprehensive cytology, Bibbo M, ed. W.B. Saunders Company, Philadelphia, 946-964 (1991).

[5] Acharya G, Mohanty PK, The morphometrical characterisation of normal blood cells of two airbreathing fishes, *Clarias batrachus* and *Anabas testudineus*. *International Research Journal of Biological Sciences* 3(11) (2014) 37-41

[6] Pala E, Dey S, Microscopy and microanalysis of blood in a snake head fish, Channa gachua exposed to environmental pollution. *Microscopy and Microanalysis* 22(1) (2016) 39-47

[7] Shen Y, Wang D, Zhao J, Chen X, Fish red blood cells express immune genes and responses. *Aquaculture and Fisheries* 3(1) (2018) 14-21

[8] Wahlby C, Sintorn IM, Erlandsson F, Borgefors G, Bengtsson E, Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *Journal of Microscopy* 215 (2004) 67-76

[9] Carpenter AE, Jones TR, Lamprecht MR, et al., CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* 7 (2006) R100

[10] Lamprecht MR, Sabatini DM, Carpenter AE, CellProfiler: Free, versatile software for automated biological image analysis. *Biotechniques* 42 (2007) 71-75

[11] Jones TR, Carpenter AE, Lamprecht MR, et al., Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences USA* 106(6) (2009) 1826-1831

[12] Ljosa V, Carpenter AE, Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening. *PLOS Computational Biology* 5(12) (2009) e1000603

[13] Kamentsky L, Jones TR, Fraser A, et al., Improved structure, function, and compatibility for CellProfiler: Modular high-throughput image analysis software. *Bioinformatics* 27 (2011) 1179-1180

[14] Bray M-A, Vokes MS Carpenter AE, Using CellProfiler for automatic identification and measurement of biological objects in images. *Current Protocols in Molecular Biology* 109 (2015) 14.17.1-14.17.13

[15] Talapatra SN, Mitra P, Swarnakar S, Morphology and phenotype of peripheral erythrocytes of fish: A rapid screening of images by using software. *International Letters of Natural Sciences* 54 (2016) 27-41

[16] LeCun Y, Bengio Y, Hinton G, Deep learning. *Nature* 521(7553) (2015) 436-444

[17] Mishra S, Dash A, Jena L, Use of deep learning for disease detection and diagnosis. In: Bhoi A, Mallick P, Liu CM, Balas V, eds. Bio-inspired Neurocomputing. Studies in Computational Intelligence. 903 Springer: Singapore 2021

[18] Bhuvaneswari E, Sarma Dhulipala VR, The study and analysis of classification algorithm for animal kingdom dataset. *Information Engineering* 2(1) (2013) 6-13

[19] Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D, Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *BioMed Research International* 2014 (2014)

[20] Hamid AJ, Ahmed TM, Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal* 3(1) (2016) 1-9

[21] Bhatia S, Sharma P, Burman R, Hazari S, Hande R, Credit scoring using machine learning techniques. *International Journal of Computer Applications* 161(11) (2017) 1-4

[22] Chakraborty I, Choudhury A, Banerjee TS, Artificial intelligence in biological data. *Journal of Information Technology & Software Engineering* 7(4) (2017) 1000207

[23] Almryad AS, Kutucu H, Automatic identification for field butterflies by convolutional neural networks. *Engineering Science and Technology, An International Journal* 23(1) (2020) 189-195

[24] Attwal KPS, Dhiman AS, Exploring data mining tool - Weka and using Weka to build and evaluate predictive models. *Advances and Applications in Mathematical Sciences* 19(6) (2020) 451-469

[25] Mishra S, Mallick PK, Tripathy HK, Bhoi AK, González-Briones A, Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. *Applied Sciences* 10(22) (2020) 8137

[26] Frank E, Hall MA, Witten IH, The WEKA workbench, Online appendix for data mining: Practical machine learning tools and techniques. Morgan Kaufmann, 4th edition 2016

[27] Bouckaert RR, Frank E, Hall M, et al., WEKA manual for version 3-8-5. University of Waikato, Hamilton, New Zealand, 2020, December 21

[28] Jamil LS, Data analysis based on data mining algorithms using WEKA workbench. *International Journal of Engineering Sciences & Research Technology* 5(8) (2016) 262-267

[29] Zainudin Z, Shamsuddin SM, Hasan S, Deep learning for image processing in WEKA environment. *International Journal of Advances in Soft Computing and its Applications* 11(1) (2019) 1-21

[30] Siraj-Ud-Doulah M, Application of machine learning algorithms in bioinformatics. *Bioinformatics & Proteomics Open Access Journal* 3(1) (2019) 000127

[31] Asaoka R, Hirasawa K, Iwase A, et al., Validating the usefulness of the “random forests” classifier to diagnose early glaucoma with optical coherence tomography. *American Journal of Ophthalmology* 174 (2017) 95-103

[32] An G, Omodaka K, Hashimoto K, et al., Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. *Journal of Healthcare Engineering* 2019 (2019)

[33] Meeradevi Sindhu N, Mundada MR, Machine learning in agriculture application: Algorithms and techniques. *International Journal of Innovative Technology and Exploring Engineering* 9(6) (2020) 1140-1146