Indian J. Genet. Plant Breed., (2024); 84(1): 92-98
https://doi.org/10.31742/ISGPB.84.1.8
ISSN: 0975-6906
www.isgpb.org

RESEARCH ARTICLE

Evaluation of resampling techniques for artificial neural network based identification of promising genotypes in sugarcane (*Saccharum officinarum* L.) varietal trials

Syed Sarfaraz Hasan, Arun Baitha[1], Lal Singh Gangwar and Sanjeev Kumar[2*]

Abstract

Identifying promising genotypes in varietal trials is one of many agriculture domain applications requiring an artificial neural network (ANN) implementation for intelligent decisions. However, varietal trial data for identification is usually imbalanced, posing challenges for neural network classification tasks. For example, only 33 genotypes were identified as promising in zonal varietal trials of AICRP on Sugarcane during 2016-21, against a non-promising class of 148. A neural network trained using the imbalanced class dataset tend to exhibit prediction accuracy according to the highest class of the dataset. Resampling techniques adjust the ratio between different classes, making the data more balanced. The study evaluated four resampling techniques viz. random under-sampling, random oversampling, and ensemble, SMOTE to balance varietal trial dataset to build ANN to identify promising genotypes in sugarcane. The paper describes the methodology used for building such a model using resampling techniques and then presents these approaches' comparative performance in identifying promising genotypes. Results indicate that SMOTE and random oversampling performed well for balancing datasets for developing neural network model in comparison to no-resampling of imbalanced datasets. SMOTE outperformed all resampling techniques by achieving high precision, recall and F1 score values for both positive and negative classes. However, ensemble and random under-sampling methods did not show good results compared to SMOTE and random over-sampling. Study will be useful in developing artificial intelligence-based tools to identify promising genotypes in varietal trials of sugarcane in particular and other crops in general.

Keywords: Artificial neural network, resampling, sugarcane, varietal trial, machine learning.

Introduction

Sugarcane is a major cash crop and raw input for the country's second-largest agro-based sugar industry. It holds 3% of gross cropped area of the country with production and productivity 411 mt and 81.5t/ha, respectively (Shukla et al. 2018). However, to contribute considerably in doubling farmers' income, the quality and productivity of cane must be enhanced. United Nations has also set the target of 50% additional production of food by 2050 from less available arable lands under changing climate, food, demography and increasing population. Estimates says that the demand for sugarcane production in India by the end of 2050 will be 630mt without increasing area requiring 105t/ha productivity.

To address these requirements, varietal improvement programs are essential to the sugarcane research system, aiming to improve production and productivity requirements and deter biotic and abiotic stresses. Identification of promising genotypes of sugarcane undergoes location-based and multiphase testing for both plant and ratoon crop under these programs. In India, multi-location trials of sugarcane genotypes are conducted by coordinated efforts of agricultural universities, research institutes, and private

Agricultural Knowledge Management Unit, [1]Division of Crop Protection, [2]Division of Crop Improvement, ICAR-Indian Institute of Sugarcane Research, Raebareli Road, PO Dilkusha, Lucknow 226 002, India.

***Corresponding Author:** Sanjeev Kumar, Division of Crop Improvement, ICAR-Indian Institute of Sugarcane Research, Raebareli Road, PO Dilkusha 226 002, Lucknow, India, E-Mail:skiisr@rediffmail.com

How to cite this article: Hasan S.S., Baitha A., Gangwar L.S. and Kumar S. 2024. Evaluation of resampling techniques for artificial neural network based identification of promising genotypes in sugarcane varietal trials. Indian J. Genet. Plant Breed., **84**(1): 92-98.

Source of support: Nil

Conflict of interest: None.

Received: Aug. 2023 **Revised:** Dec. 2023 **Accepted:** Feb. 2024

sector under ambit of the All India Coordinated Research Project on Sugarcane. Trials are regularly monitored and data on more than twenty characters such as germination %, tillers, shoots, NMC, fibre, brix, sucrose, CCS, cane yield, etc. are collected frequently at different stages of crop. Promising genotypes are identified after a series of analysis and discussions on trial data and pooled data. Accurate and reliable analysis of characters to identify promising genotypes for further commercial release is important to improve the performance of the sugarcane sector. It is a quite complex and time-consuming task in traditional computing; information about some important characters may remain unnoticed by experts despite best efforts and artificial intelligence techniques need intervention.

Inspired by artificial neural network (ANN) abilities and concerns of varietal trials, we have used this technique to identify promising genotypes in sugarcane varietal trials. Technique extract higher-level features from the raw input for making intelligent decisions.Various domains including pattern recognition, computer vision, and natural language processing have witnessed the great power of neural networks. ANN approach is data driven and results depend on quality of data to learn a model from. For the development of the model, pooled data of Zonal Varietal Trials (ZVT), monitoring reports of trials and red rot evaluation data for duration 2016-21 has been taken from secondary sources of AICRP on Sugarcane. The data of 181 genotypes, proposed for multi-location trials from states of peninsular, north-west, north-central & north-eastern, and east-cost zones in this duration, was collected.

However, collected data shows huge imbalance in number of promising and non-promising genotypes in varietal trial data. There were only 33 promising genotypes found as against 148 non-promising genotypes. ANN models constructed and trained with imbalanced data cannot recognize minority data well.Model so developed will recognize majority data well but have poor performance on recognizing minority data and pose a major challenge (Bagui and Li 2021). Imbalanced data sets exist widely in real world and they have been providing great challenges for classification tasks (Wang et al. 2016). Fraud detection, churn prediction, spam detection, claim prediction, anomaly detection, and outlier detection are examples of imbalanced data. Class imbalance problem is considered one of the emerging challenges in the machine learning area (Yang and Wu 2006; He and Garcia 2009; Fernandez et al. 2011).

Leevy et al. (2018) surveyed the problem of class imbalance and found that solutions were mainly divided into data-level and algorithm-level methods.To handle an imbalanced dataset, More (2018) reviewed a number of resampling techniques, including random undersampling of the majority class, random oversampling of the minority class, SMOTE, and many other techniques. Wallace et al. (2011) used SMOTE with SVM as the base classifier while Hulse et al. (2007) demonstrated that simple undersampling tends to outperform SMOTE in low-dimensional data. Abdi and Sattar (2016) proposed a new oversampling algorithm based on Mahalanobis distance and showed how it generates less duplicate and overlapping data points as opposed to other oversampling techniques. Dong et al. (2019) designed a model based on batch-wise incremental minority (sparsely sampled) class rectification by hard sample mining in majority (frequently sampled) classes.A novel Balance Cascade-based kernelized extreme learning machine to handle the problem of class imbalance has been designed by Raghuwanshi and Shukla (2020). Cieslak et al. (2006) have used SMOTE to detect network traffic intrusions. Ertekin *et al*.(2007) and Radivojac et al. (2004) have also evaluated SMOTE's performance based on the number of samples.

Most of the studies on neural networks focus on balanced datasets, while its performance on imbalanced dataset is not well examined. One of the approaches used to deal with class imbalance problems, called data approach, consists of resampling the data in order to balance the classes before building the classifier. This approach is independent of the learning algorithm used and most of the research has been done in this direction (Berry et al. 2000; Japkowicz and Stephen, 2002; Estabrook et al. 2004). However, most of the studies conducted are for performing specific resampling techniques to solve domain problems. However, no comparative evaluation of resampling techniques have been made in these studies for ANN model development, particularly in agriculture. In this study, while developing ANN model for the prediction of promising genotypes, four resampling techniques viz. random under-sampling, random oversampling, ensemble, SMOTE, etc. has been applied and evaluated along with no-resampling. Paper describes the methodology used in our approach for building an artificial neural network model and then presents results of its evaluation under various resampling techniques.

Materials and methods

For the development and evaluation of artificial neural network, data is bifurcated in training and testing datasets, out of which the training set is used for training and building the model, while testing dataset is used for testing model. For this study, zonal varietal trial data of AICRP (Sugarcane) of duration 2016-21 was undertaken from Principal Investigator's Reports of Crop Improvement (Ram 2017, 2018, 2019, 2020, 2021) and Principal Investigator's Report of Plant Pathology (Viswanathan 2017, 2018, 2019, 2020, 2021). Six characters have been used to predict the promising genotypes, as shown in the data structure in Table 1. The last field signifies the score 1 and 0 for promising

Table 1. Structure of data used in the study in this study

Field name	Field type	Field description
Monitoring Score	Float	Average numeric score of monitoring team score for trial
Red Rot Resistance Score	Binary	Binary value of 1 to indicate genotype is red rot resistant / moderately resistant otherwise 0
Cane Yield Difference	Float	Difference of cane yield t/ha from best standard of the zone
CCS % Difference	Float	Difference of CCS% from best standard of the zone
CCS Yield Difference	Float	Difference of CCS yield t/ha from best standard of the zone
Sucrose Difference	Float	Difference of sucrose % from best standard of the zone
Identified Class	Binary	Promising genotype indicator. 1 for promising, 0 for non-promising

and non-promising classes, respectively. Model building and evaluation using resampling techniques in this study undergoes the following phases:

Dataset creation

In first phase following data management activities were performed to prepare the dataset for next phase:

Data was collected and compiled using spreadsheet and database management techniques.

Data was pre-processed to transform into a much-desired form so that useful information can be derived. Some of the activities performed in this stage were data transformation in a common unit, converting quality data into quantitative form, missing/wrong values correction, removal of undesired data, and data pooling, etc.

TensorFlow libraries under Python programming platform have been used to build dataset and model using the above data. It undergoes the following activities to build dataset for modeling:

- Imported TensorFlow and other important libraries in Python.
- Imported data created above to be used in our program.
- Scaled the data to bring it in same scaling range.
- Bifurcated datasets into positive (minority) and negative (majority) classes.

Model construction

Artificial neural network is inspired by the structure and function of the human brain, in which multiple layers of processing are used in network mode for making intelligent decisions. ANN consists of artificial neurons, mimicking a biological neuron's function, as shown in Figure 1. The output from each neuron is an activated sum of input multiplied by the weight to that neuron and the bias value.

Neural networks can learn weights that map any input to the output. An artificial neural network is the simplest form of a feed-forward neural network because inputs are processed only in the forward direction. Further, this network can learn any non-linear dataset with the help of non-linear and sigmoid functions (Wen et al. 2018). It is a network of input, output and hidden layers. The input layer consists of six input parameters with respect to data input as shown in Table 1. Output layer corresponds to two classes

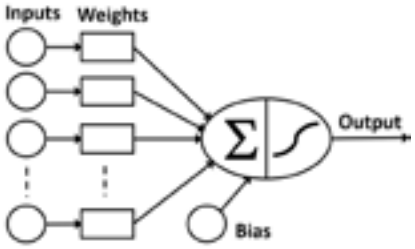

Fig. 1. Structure of artificial neural network

of promising and non-promising genotypes. Two sets of hidden layers have been used in our model, which assist in extracting features in every iteration of learning. Another important consideration while model development is hyper parameters settings. These parameters are used to tune the network for desired learning performance. Important hyperparameters configured to train the model are number of epochs, batch size and learning rate.

Resampling

Dataset obtained in this study consists of 181 entries of sugarcane genotypes, in which 33 are promising genotypes and rest 148 are non-promising. To address the imbalance issue in this dataset for neural network-based identification, following resampling techniques has been used along with no-resampling.

Random under-sampling: For balancing dataset, random samples of minority class size were picked from majority class. By merging newly drawn samples and minority class samples, we have constructed resample of size 66 for the construction of model.

Random oversampling: In contrary to random undersampling, in this case samples were randomly picked from minority class equal to size of majority class. Newly drawn and majority class samples were merged to be used as resample of size 296 for the construction of model.

Synthetic Minority Oversampling Technique (SMOTE): In this case, a point is randomly picked from the minority class and k-nearest neighbours for this point are computed. The synthetic points are added between the chosen point and its neighbours. SMOTE function from imblearn.over. sampling library has been used to resample dataset which

Table 2. List of genotypes entries and standards used

Zone and entry group	Test entries	Standard
East Coast Zone (Early)	Co 13023, CoA 12321, CoA 12322, CoA 12323, CoA 13322, CoA 13323, CoA 14321, CoA 16321, CoC 13336, CoC 13337, CoC 14336, CoC 15336, CoC 15338, CoC 16336, CoC 16337, CoOr 12346, CoV 12356, CoV 13356, CoV 15356, CoV 16356	Co 6907, CoA 92081, CoC 01061, CoOr 03151
East Coast Zone (Mid-late)	Co 13028, Co 13029, Co 13031, CoA 11326, CoA 12324, CoA 14323, CoC 13339, CoC 14337, CoC 15339, CoC 16338, CoC 16339, CoOr 13346, CoOr 15346, CoV 16357, PI 14377	Co 06030, Co 86249, CoV 92102
North Central Zone + North Eastern Zone (Early)	CoLk 12207, CoLk 14206, CoLk 15466, CoLk 15467, CoP 11436, CoP 11437, CoP 11438, CoP 12436, CoP 13437, CoP 14437, CoP 15436, CoSe 11451, CoSe 12451, CoSe 13451, CoSe 13452, CoSe 14451, CoSe 14454, CoSe 15452, CoSe 15455	BO 130, CoLk 94184, CoSe 01421, CoSe 95422
North Central Zone + North Eastern Zone (Mid-late)	BO 155, CoLk 09204, CoLk 12209, CoLk 14208, CoLk 14209, CoLk 15468, CoLk 15469, CoP 12438, CoP 14438, CoP 14439, CoP 15438, CoP 15439, CoP 15440, CoSe 11453, CoSe 11454, CoSe 11455, CoSe 12453, CoSe 14455, CoSe 15453, CoSe 15454	BO 91, CoP 06436, CoP 9301, CoSe 92423
North West Zone (Early)	Co 13034, Co 14034, Co 15023, Co 15024, Co 15027, CoH 11262, CoLk 11201, CoLk 11202, CoLk 11203, CoLk 14201, CoLk 15201, CoLk 15205, CoPb 13181, CoPb 14181, CoPb 14211, CoPb 15212, CoS 13231	Co 0238, Co 05009, CoJ 64
North West Zone (Mid-late)	Co 11027, Co 12029, Co 13035, Co 14035, CoH 11263, CoH 12263, CoH 13263, CoH 14261, CoLk 11204, CoLk 11206, CoLk 12205, CoLk 13204, CoLk 14203, CoLk 14204, CoPant 12226, CoPant 13224, CoPb 11214, CoPb 12211, CoPb 13182, CoPb 14184, CoPb 14185, CoS 11232, CoS 12232, CoS 14233	Co 05011, CoPant 97222, CoS 767, CoS 8436
Peninsular Zone (Early + Midlate)	Co 09009, Co 10004, Co 10005, Co 10006, Co 10015, Co 10017, Co 10024, Co 10026, Co 10027, Co 10031, Co 10033, Co 11001, Co 11004, Co 11005, Co 11007, Co 11012, Co 11019, Co 12007, Co 12008, Co 12009, Co 12012, Co 12019, Co 12024, Co 13002, Co 13003, Co 13004, Co 13006, Co 13008, Co 13009, Co 13013, Co 13014, Co 13018, Co 13020, Co 14002, Co 14004, Co 14012, Co 14016, Co 14027, Co 14030, Co 14032, CoM 10083, CoM 11081, CoM 11082, CoM 11084, CoM 11085, CoM 11086, CoM 12085, CoN 13072, CoN 13073, CoN 14073, CoSnk 13101, CoSnk 13103, CoSnk 13106, CoSnk 14102, CoSnk 14103, CoT 10366, CoT 10367, CoT 10368, CoT 10369, CoT 14367, CoTl 14111, CoVC 10061, CoVC 14062, MS 13081, MS 14081, MS 14082, PI 10131, PI 10132, PI 13132, VSI 12121	Co 85004, Co 86032, Co 94008, Co 99004, CoC 671, CoSnk 05103

produced 296 records.

Ensemble with undersampling: In ensemble method, the number of samples formed using scale consists of a ratio of the min class number and max class number. Using ensemble technique total 5 samples with sample size of 66(for 4 samples) and 42 (for 1 sample) were formed. Models were constructed for all these groups and majority decision was undertaken from 5 groups to classify the test data.

No-resampling: No resampling was performed and thus dataset consist of 33 promising genotypes (minority class) and 148 non-promising (majority class).

Model evaluation

ANN models were developed using resampling and no-resampling techniques mentioned above using TensorFlow library in Python language. Dataset created in phase one was first resampled using resampling techniques mentioned above and then bifurcated in training and test set in ratio of 4:1 using NumPy libraries along with Python codes. Sequential models were constructed and trained using training data and then tested using test data. Confusion matrix helped in identifying True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) cases out of the model. Metrics viz. accuracy, precision, recall, F1 score has been used to evaluate the model. Accuracy is a metric to describe the performance of model across all classes. It is calculated as the number of correct predictions made by the model to the total number of predictions. The precision metric tells what percentage of all the positively predicted classes are actually positive classes. Recall measures out of all the positive classes and what percentage of them were actually predicted as positive. Recall metric quantifies the number of correct positive predictions made out of all positive (promising) predictions that could have been made. F1 score is harmonic mean of Precision and Recall. F1 measure provides a way to combine both precision and recall into a single measure that captures both properties. Average values of these metrics were calculated, and evaluated corresponding to a range of epochs for resampling techniques using formulae given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 3. Sample of dataset developed in first phase

Monitoring score	Red rot resistance score	Cane yield difference	CCS % Difference	CCS Yield Difference	Sucrose Difference	Identified Class
2.566406	1	3.759766	-0.640137	0.509766	-0.77002	0
1.527344	1	-3.429688	-0.340088	-0.839844	-0.540039	0
3.599609	1	4.539062	0.340088	-0.27002	-0.429932	0
2.208984	0	-1.360352	-0.620117	-0.850098	-0.810059	0
2.482422	1	5.871094	0.360107	1.160156	0.379883	1
3.533203	1	3.080078	0.059998	-0.419922	-0.589844	0
1.348633	1	-8.171875	-0.130005	-1.080078	-0.219971	0
2	1	3.179688	-0.180054	-0.209961	-0.040009	1
2.333984	1	-9.476562	-1.19043	-2.480469	-1.549805	0
1.733398	1	12.03125	-0.689941	1.049805	-0.680176	0
2.541016	1	3.949219	-0.080017	0.669922	-0.150024	0
3.849609	1	-5.488281	-0.150024	-0.209961	-0.310059	0
2	1	-2.859375	-0.080017	-0.429932	-0.130005	0
2.523438	1	-11.078125	-0.25	-1.540039	-0.409912	0
2.259766	1	-2.089844	-0.469971	-0.680176	-0.680176	0
2.566406	1	3.759766	-0.640137	0.509766	-0.77002	0
1.527344	1	-3.429688	-0.340088	-0.839844	-0.540039	0
3.599609	1	4.539062	0.340088	-0.27002	-0.429932	0

Results and discussion

Dataset in this study consists of 181 sugarcane genotypes proposed for multi-location trials during 2016-21 in AICRP (Sugarcane) zonal varietal trials. Table 2 shows list of entries used of early and mid-late groups along with standards for the north-west, north-central & north-eastern, east-cost, and peninsular zones. In this dataset of 33 genotypes has been identified by AICRP (Sugarcane) as promising and rest 148 as non-promising in zonal varietal trials as shown in table. Crop characters covered are cane yield, CCS%, CCS yield and sucrose % along with monitoring and red rot scores.

Data was pre-processed, scaled and then resampled using various techniques. Table 3 shows the sample processed data used by the system. First six columns in this table correspond to monitoring score, red rot resistance score and four crop characters (cane yield, CCS %, CCS yield, sucrose %) difference from standards, while last column signifies the score 1 and 0 for promising and non-promising classes respectively.

Figure 2 shows neural network comprising of input, output and hidden layers. Input layer in this model consists of six neurons with respect to input parameters viz. monitoring score, red rot resistance score, and difference of cane yield, CCS yield, CCS % and sucrose % from standards. Output layer corresponds to two classes of promising and non-promising genotypes represented by binary output of 1 and 0. Two hidden layers have been used in this model, which assist in extracting features in every iteration of learning.

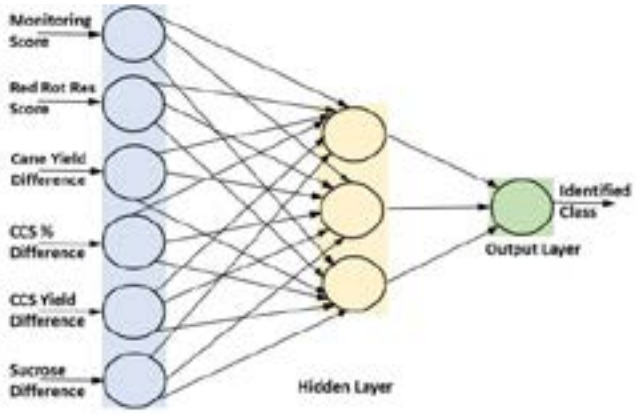

Fig. 2. Multi-layer neural network model used

Accuracy, precision, recall and F1 score were measured to evaluate models built using various resampling techniques compared to no-resampling. Table 4 shows a comparison of these parameters for the positive class(minority class representing promising genotypes), and negative class (majority class representing non-promising genotypes).

A model built with no resampling shows the highest accuracy of 0.85 compared to all deployed resampling techniques. It can be seen that precision, recall and F1 score values for the negative class are quite high compared to the positive class due to a huge imbalance in a dataset. High accuracy in this case is clearly influenced by big size of the negative class and therefore, precision and recall values of positive class predictions are quite low.

Table 4. Evaluation metrics corresponding to resampling techniques

Resampling Technique	Dataset size	Accuracy	Positive Class			Negative Class		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
SMOTE	296 (Majority: 148, Minority: 148)	0.80	0.81	0.78	0.79	0.79	0.82	0.80
Random Oversampling	296 (Majority: 148, Minority: 148)	0.78	0.78	0.76	0.76	0.76	0.80	0.77
Ensemble with Undersampling	66 (Majority: 33, Minority: 33) for 4 group and 42 (Majority:21, Minority: 21) for 1 group	0.65	0.35	0.77	0.47	0.84	0.62	0.71
Random Undersampling	66 (Majority: 33, Minority: 33)	0.64	0.63	0.72	0.66	0.58	0.55	0.55
No Resampling	181 (Majority: 148, Minority: 33)	0.85	0.49	0.40	0.42	0.88	0.95	0.91

SMOTE and random oversampling resampling techniques showed quite high and equivalent values of precision, recall and F1 score metrics for both positive and negative classes, although overall accuracy shown by these models is a little low in comparison to no-re-sampling. The random oversampling technique achieved equivalent values of F1 score metric for positive and negative classes as 0.76 and 0.77, respectively, with an accuracy of 0.78. This indicates that the random oversampling technique performed well for balancing imbalanced datasets to develop an artificial neural network model for selecting promising genotypes. However, SMOTE outperformed all resampling techniques by achieving high precision, recall and F1 score values for both positive and negative classes. This technique achieved the highest accuracy 0.80 among resampling techniques as also achieved highest equivalent F1 score of 0.79 and 0.80 for positive and negative classes, respectively, compared to other resampling techniques adopted.

Ensemble and random under-sampling methods did not show good results in terms of all metrics compared to SMOTE and random over-sampling techniques. These methods showed an accuracy of only 0.64 to 0.65 with F1 score in the range of 0.47 to 0.71, which are quite low for the reliability of the model for the prediction of promising genotypes. Differences between F1 scores for positive and negative classes are quite high, indicating that these techniques didn't solve class imbalance problems well compared to SMOTE and random over-sampling.

Researchers have proposed several approaches to deal with imbalanced datasets and improve the classifiers' quality. Major ways to manage the imbalanced classes in the dataset are changing performance metrics, adding more data, experimenting with different algorithms, and resampling the dataset. Results reported in this work may be used further for the selection of correct resampling techniques for imbalanced datasets, particularly for agriculture domain problems. It will improve the accuracy and reliability of ANN models developed in such scenarios. However, further studies will be beneficial for evaluating these techniques while dealing with big data in agriculture. Studies conducted will aid in developing artificial intelligence-based tools for the automatic identification of promising genotypes in varietal sugarcane trials. Further, these results will also be beneficial in the artificial neural network-based identification of promising genotypes in varietal trials of other crops.

Authors' contribution

Conceptualization of research (SK, SSH); Designing of the experiments (SK, SSH, LSG); Contribution of experimental materials (SK, LSG); Execution of field/lab experiments and data collection (SSH, AB); Analysis of data and interpretation (SK, SSH, LSG); Preparation of manuscript (SK, SSH, LSG, AB).

Acknowledgment

We acknowledge the Project Coordinator, All India Coordinated Research Project on Sugarcane for the reports and data used in this study.

References

Abdi L. and Sattar H. 2016. To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans. Knowl. Data Eng., **28**(1): 238–51.

Bagui S. and Li K. 2021. Resampling imbalanced data for network intrusion detection datasets. J. Big Data, **8**(6): https://doi.org/10.1186/s40537-020-00390-x.

Berry M.A. and Linoff, G.S. 2000. "Mastering Data Mining: The Art and Science of Customer Relationship Management", Industrial Management & Data Systems, **100**(5): 245-246. https://doi.org/10.1108/imds.2000.100.5.245.2

Chen H., Yining L., Chen C.L. and Xiaoou T. 2016. Learning Deep Representation for Imbalanced Classification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5375-5384.

Cieslak D.A., Chawla N.W. and Striegel A. 2006. Combating Imbalance in Network Intrusion Datasets. Proc. IEEE Int. Conf. Granular Computing, Atlanta, Georgia, USA, pp. 732-737.

Dong Q., Gong S. and Zhu X. 2019. Imbalanced Deep Learning by Minority Class Incremental Rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, **41**(6): 1367-1381. doi: 10.1109/TPAMI.2018.2832629.

Ertekin S.E., Huang J., Bottou L. and Giles C.L. 2007. Learning on the border: active learning in imbalanced data classification.

Proc. ACM Conference on information and knowledge management, Lisbon, Portugal, pp. 127–36.

Estabrooks A., Jo T.J. and Japkowicz N. 2004. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, **20**(1): 18–36.

Fernández A., García S. and Herrera F. 2011. Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. In Corchado E, Kurzyński M. and Woźniak M., eds., HAIS 2011, Part I. LNCS, **6678**: 1–10.

He H. and Garcia E. 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng., **21**(9): 1263–1284.

Hulse J.V., Khoshgoftaar T.M. and Napolitano A. 2007. Experimental perspectives on learning from imbalanced data. Proc. 24th international conference on machine learning, Corvallis, Oregon: Oregon State University,pp. 935–42.

Japkowicz N. and Stephen S. 2002. The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis J., **6**(5): 429–449.

Leevy J.L., Khoshgoftaar T.M., Bauder R.A. and Seliya N. 2018. A survey on addressing high-class imbalance in big data. J. Big Data, **5**: 42. https://doi.org/10.1186/s40537-018-0151-6.

Mohri M., Rostamizadeh A. and Talwalkar A. 2018. Foundations of machine learning. 2nd ed. Cambridge: MIT Press, 2018.

More A. 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets.*ArXiv*.doi: 10.48550/arXiv.1608.06048.

Radivojac P., Chawla N.V., Dunker A.K. and Obradovic Z. 2004. Classification and knowledge discovery in protein databases. J. Biomed. Inform., **37**(4): 224–39. https://doi.org/10.1016/j.jbi.2004.07.008.

Raghuwanshi B.S. and Shukla S. 2020. SMOTE based class-specific extreme learning machine for imbalanced learning. Pattern Anal. Appl., **187**: 104814.

Ram B. 2017. Principal Investigator's Report 2016-17. Varietal Improvement Programme, All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-691. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PI Report-CI-2016-17.pdf.

Ram B. 2018. Principal Investigator's Report 2017-18. Varietal Improvement Programme, All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-706. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PI Report-CI-2017-18.pdf.

Ram B. 2019. Principal Investigator's Report 2018-19. Varietal Improvement Programme, All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-504. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PI Report-CI-2018-19.pdf.

Ram B. 2020. Principal Investigator's Report 2019-20. Varietal Improvement Programme, All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-554. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PI Report-CI-2019-20.pdf.

Ram B. 2021. Principal Investigator's Report 2020-21. Varietal Improvement Programme, All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-579. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PI Report-CI-2020-21.pdf.

Shukla S.K., Yadav S.K. and Pathak A.D. 2018. Low cost technologies in sugarcane agriculture. ICAR–All India Coordinated Research Project on Sugarcane, IISR, Lucknow. pp. 1-55.

Viswanathan R. 2017. Technical Report, Plant Pathology (2016-17). All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-117. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PIReport-PP-2016-17.pdf.

Viswanathan R. 2018. Technical Report, Plant Pathology (2017-18). All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-133. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PIReport-PP-2017-18.pdf.

Viswanathan R. 2019. Technical Report, Plant Pathology (2018-19). All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-130. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PIReport-PP-2018-19.pdf.

Viswanathan R. 2020. Technical Report, Plant Pathology (2019-20). All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-123. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PIReport-PP-2019-20.pdf.

Viswanathan R. 2021. Technical Report, Plant Pathology (2020-21). All India Coordinated Research Project on Sugarcane. ICAR-Sugarcane Breeding Institute, Coimbatore, 1-131. Available at https://iisr.icar.gov.in/iisr/aicrp/download/PIReport-PP-2020-21.pdf.

Wallace B., Small K., Brodley C. and Trikalinos T. 2011. Class imbalance, redux. Proc. IEEE 11th international conference on data mining (ICDM), Vancouver, Canada, pp. 754–63.

Wang S., Liu W., Wu J., Cao L., Meng Q. and Kennedy P.J. 2016. Training deep neural networks on imbalanced data sets.Proc. International Joint Conference on Neural Networks (IJCNN), pp. 4368-4374. doi: 10.1109/IJCNN.2016.7727770.

Wen M., Cong P., Zhang Z., Lu H. and Li T. 2018. DeepMirTar: a deep learning approach for predicting human miRNA targets. Bioinformatics, **34**(22): 3781–87.

Yang Q.and WuX. 2006. 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, **5**(4): 597–604.

Yun Q., Yanchun L., Mu L., Guoxiang F. and Xiaohu S. 2014. A resampling ensemble algorithm for classification of imbalance problems. Neurocomputing, **143**: 57-67. https://doi.org/10.1016/j.neucom.2014.06.021.