Indonesian Journal of Electrical Engineering and Computer Science
Vol. 28, No. 2, November 2022, pp. 926~933
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v28.i2.pp926-933

Machine learning ensemble approach for healthcare data analytics

Deepali Pankaj Javale, Sharmishta Suhas Desai
School of Computer Engineering and Technology, MIT World Peace University, Pune, India

Article Info

Article history:

Received Apr 1, 2022
Revised Jul 4, 2022
Accepted Jul 28, 2022

Keywords:

Adaptive synthetic
Ensemble learning
Healthcare
Stacking-C
Synthetic minority oversampling technique

ABSTRACT

In healthcare machine learning is used mainly for disease diagnosis or acute condition detection based on patient data analysis. In the proposed work diabetic patient dataset analysis is done for hypoglycemia detection which means the lowering of blood glucose level (BGL). Often in healthcare it is observed that the dataset is imbalanced. Therefore, an Ensemble Approach using imbalanced dataset techniques synthetic minority over-sampling technique and adaptive synthetic oversampling methods with different evaluation methods like train-test, K-fold, stratified K-Fold and repeat train-test were used. This ensemble approach was implemented on diabetic dataset using K-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), Naïve Bayes (NB) and logistic regression classifiers with average Stacking-C method thereafter to conclude. Comparative analysis was done using three different considerations. The results showed that KNN and random forest gives more stable metric values both on balanced and imbalanced dataset. The confusion matrix consideration concluded that KNN and random forest were found to be better with least false negative and maximum true positive count. But if average train and test time is taken into consideration then Naïve Bayes and random forest had least average train-test time. Thus, the three different considerations concluded that the proposed ensemble approach gives better clarity for different classifier implementation using machine learning.

Corresponding Author:

Deepali Pankaj Javale
School of Comuter Engineering and Technology, Faculty of Engineering, MIT World Peace University
Pune, India
Email: deepali.javale@mitwpu.edu.in

1. INTRODUCTION

It is often observed that in healthcare dataset binary classification majority of the dataset are imbalanced. Especially in healthcare dataset the critical condition records are significantly less as compared to normal condition records. The critical conditions like heart attack condition, asthama attack and hypoglycemia conditions occur very rarely. Thus giving rise to highly imbalanced dataset. The imbalanced dataset handling if not taken into consideration and if not balanced then it may often lead to wrong or inappropriate conclusions. For this dataset balancing plays very critical and important role. Also as compared to only classifier base comparative analysis ensemble technique is always a preferred one.

Diabetes is considered one of fastest spreading chronic disease. The experimentation work was carried out on diabetes dataset for hypoglycemia detection. Hypoglycemia means the lowering of blood glucose level which may lead to severe complications for an individual [1]. We have made and attempt for analyzing detection of hypoglycemia using superficial body parameter values. Usually it is observed that to get the blood glucose level (BGL) value often a pricking method is used. An attempt was made to find out

hypoglycemia occurrence in diabetes people using their superficial body parameters [2], [3]. For any type of dataset for accurate decision-making machine learning has played and important role. Various classifiers and the performance metrics available have given the researcher a good scope for healthcare analysis [4]-[6]. The literature survey is mainly focused on imbalanced dataset handling, evaluation methods using machine learning and ensemble techniques. In one of the paper the authors mainly focused on handling the imbalanced dataset problem. The SVM-SMOTE method was used for dataset balancing. A hybrid model implementation was done with KFold cross validation with K=5 for evaluating the classifier performance. A hybrid model using genetic algorithm-based feature selection with stacking method of ensemble technique was used [7].

In one of the related work the author has used two oversampling techniques for dataset imbalance viz. Synthetic minority oversampling technique (SMOTE) and adaptive synthetic (ADASYN). The pima dataset was considered for implementation with 4 different classifiers random forest, logistic regression, XA boost and support vector machine (SVM). A significant growth in the metric values was reflected after using SMOTE and ADASYN oversampling technique for imbalanced dataset [8]. Binary classification using imbalanced and balanced techniques was discussed. ensemble of classifiers based on multiobjective genetic sampling for imbalanced classification (E-MOSAIC) method was elaborated. This was a genetic algorithm based innovative way to avoid the loss of information due to under sampling and to avoid repeated information due to oversampling [9]. Iimbalanced dataset often lead to lowered performance. The authors considered an intrusion detection system datasets for machine learning analysis. 6 different IDS were considered where high imbalanced dataset was observed. SMOTE was used for balancing with improved performance [10].

A novel way of first preprocessing the dataset for imbalance and converting to balanced dataset is proposed. Then this balanced dataset was given as training dataset to ensemble classifier model. Classifier ensemble hybrid approach was used with two phases viz. Resampling dataset using SMOTE for balancing and StackingC classifier ensemble. The general classifier approach was then compared with the hybrid approach proposed which showed significant rise in AUC score [11]. Also an unique ensemble strategy for medical diagnosis was given to SMOTE by using cross validation technique. In the last phase weighted majority voting strategy was used to prove the efficiency of the ensemble proposed [8]. Further the study was to explore more on performance metrics in machine learning. Just relying on accuracy score was not a good choice. In one of the article authors have given the comparative study of different metrics used in machine learning for imbalanced dataset. The difference in majority and minority class affects the metrics like accuracy and F1-score, while s area under the receiver operating characteristic curve metric shows no affect [12]. Different ensemble approach for machine learning viz. bagging, breiman boosting, and freund boosting. Imbalanced dataset are mainly to be taken into consideration. Different metrics for imbalanced dataset were discussed and experimented with. AUC was considered to be most robust [13]. Finally talking about the evaluation methods, cross validation method proves to give good results. Cross validation method is often a good choice when ambiguity in selecting train and test dataset arises. Details discussion on cross validation techniques was done on different dataset based classifier implementations. The cross validation using KFold and Stratified KFold was implemented and compared with without cross validation classifier metric values. Different K value evaluation was done [14].

2. METHOD

The very first step is dataset generation. Dataset generation in this case was the real time data inputs taken from 13 different patients. The details for dataset are shared below. The mentioned dataset is also available on IEEE data port [15].

2.1. Dataset

Data collection was done from diabetic patients in real time for the features selected. 13 different patients of different age groups recorded readings using calibrated wearable devices and apparatus. The continuous glucose monitoring kit named Freestyle librePro sensor was used to get the blood glucose level readings. While the rest of the parameter readings was taken from riversong wave O2 colored smart band. The Librepro device and riversong wave bands were calibrated together for time settings [15].

Dataset Features: Dataset with around 70000 record [15] was ready having following features,

a) Including diabetic and non diabetic people record.
b) Structured dataset with features used for experimentation as Diastolic BP, Systolic BP, Heart Rate, Shivering, Body Temperature, age, hypoglycemia detected and prehypoglycemia.
c) While dataset fields like blood glucose level (BGL), SPO2, Sweating and diabetic/nondiabetic were not used.
d) Hypoglycemia was used as target field.

– **Indian Copyright Filled** - Diary Number: 15042/2021-CO/L
Title- Dataset for people for their blood glucose level with their superficial body feature readings.

2.2. Ensemble approach used

It is often observed that healthcare dataset is usually imbalanced. The number of records for disease diagnosed is less than the number of records where disease is not diagnosed. Therefore, an Ensemble approach was used for Hypoglycemia detection analysis using superficial body parameter readings [16], [17].

2.3. Implementation

Ensemble approach implemented was a three step method given in Figure 1.
Step 1: Using Oversampling method for converting imbalanced dataset to balanced dataset [15]. Dataset balancing was done by oversampling technique SMOTE [16] and ADASYN [17]-[20]. SMOTE The original imbalanced dataset had total 70943 records. The major imbalance was found in original dataset viz. imbalance for hypoglycemia detected field. The original dataset had following imbalance count Total record: 70943, Hypoglycemia detected count records: 9055, Hypoglycemia not detected count records: 61888.

Step 2: Evaluation done by 4 different ways viz. crossfold validation, stratified crossfold validation, train-test and repeat train test method [21], [22].
Cross Validation: The cross-validation technique for above 5 models using machine learning was done with KFold and Stratified KFold method with K value being 10 and 20. The stratified KFold is considered to be one of the ways to handle dataset imbalance.
Train Test: The train test variation is done by maintaining the train: Test ratio to 7:3 and 8:2. The repeat train test strategy is adopted to get more accurate results. The repeat train test value was set to 10 and 20 and then classifier evaluation was done for 7:3 and 8:2 ratios.

Step 3: Average StackingC strategy was used at last stage to come to conclusion.

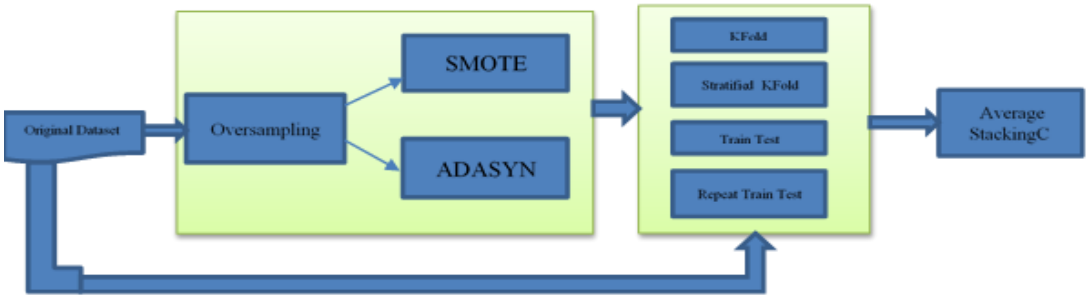


Figure 1. Ensemble approach for hypoglycemia detection

The ensemble approach with cross validation and train test technique was used for following five supervised model or classifier [23], [24] implementations viz. K-nearest neighbor (KNN), SVM, random forest, Naïve Bayes and logistic regression. The cross validation technique makes the result more concrete. Also the train test method with repeat with every time different combination will remove the possibility of skipping any important record.

2.4. Comparative analysis

Comparative analysis is done for all above datasets with all different experimentation strategies. Machine learning experimentation was done based considering metrics [25]-[27] AUC score, accuracy, F1 score, precision, recall, train time and test time. As per the literatures studied [12]-[14] ROC_AUC_Score is a robust metric for imbalanced dataset as compared to other metrics like accuracy_score, F1-Score, precision and recall.

3. RESULTS AND DISCUSSION

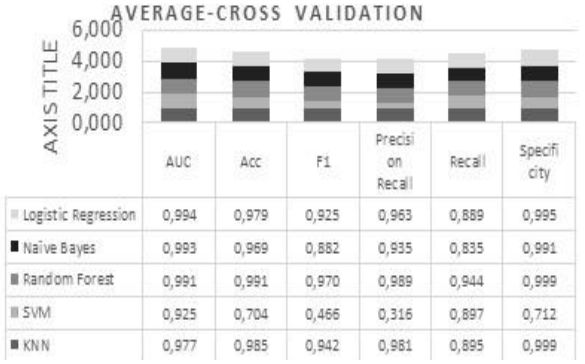
The two main categories for experimentation were imbalanced and balanced dataset. The target feature is fixed to Hypoglycemia detected which is binary and which indicated Hypoglycemia detected is 1

and vice versa. Different dataset and strategies were used for result generation. Dataset used were either original imbalanced dataset or balanced dataset by SMOTE or ADASYN method. Strategy can be either of cross validation, stratified crossfold, train-test or repeat train-test.

3.1. Imbalanced dataset experimentation

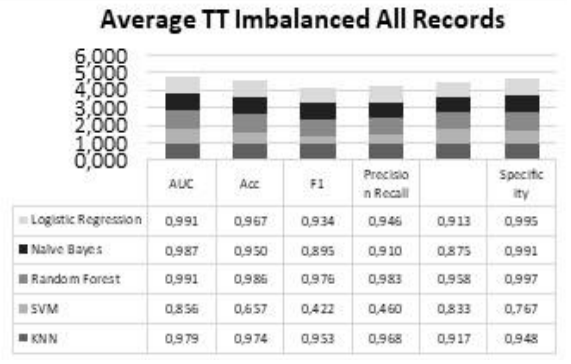
The 4 results shown in Figure 2 and 3 shows that KNN, random forest, Naïve Bayes and logistic regression are the four classifiers which gives good metric values for imbalanced dataset and AUC score is more of a balanced metric. The balancing was done for hypoglycemia detected record imbalance. Total records in the dataset were 70.943 in which hypoglycemia detected were 9.055 records and remaining 61.888 were no hypoglycemia detected records. Which was a severe imbalance. Therefore oversampling was done on the original imbalanced dataset using SMOTE and ADASYN method.

Target – Hypoglycemia Detection
Dataset –Original imbalanced dataset
Strategy – Cross Validation



	AUC	Acc	F1	Precision Recall	Recall	Specificity
Logistic Regression	0,994	0,979	0,925	0,963	0,889	0,995
Naïve Bayes	0,993	0,969	0,882	0,935	0,835	0,991
Random Forest	0,991	0,991	0,970	0,989	0,944	0,999
SVM	0,925	0,704	0,466	0,316	0,897	0,712
KNN	0,977	0,985	0,942	0,981	0,895	0,999

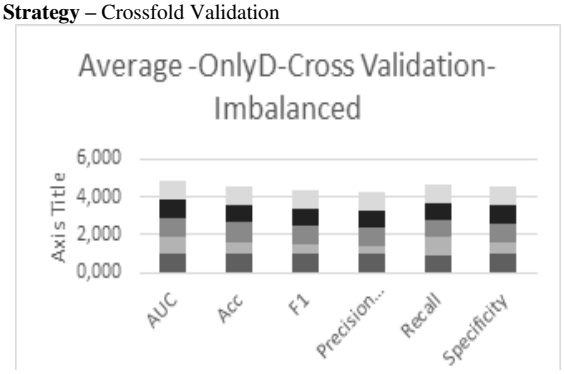
Target – Hypoglycemia Detection
Dataset –Original imbalanced dataset
Strategy – Train Test



	AUC	Acc	F1	Precision Recall		Specificity
Logistic Regression	0,991	0,967	0,934	0,946	0,913	0,995
Naïve Bayes	0,987	0,950	0,895	0,910	0,875	0,991
Random Forest	0,991	0,986	0,976	0,983	0,958	0,997
SVM	0,856	0,657	0,422	0,460	0,833	0,767
KNN	0,979	0,974	0,953	0,968	0,917	0,948

Figure 2. Imbalanced dataset experiment analysis

Target – Hypoglycemia Detection
Dataset –Only Diabetic imbalanced dataset
Strategy – Crossfold Validation



Target – Hypoglycemia Detection
Dataset –Only Diabetic imbalanced dataset
Strategy – Train Test

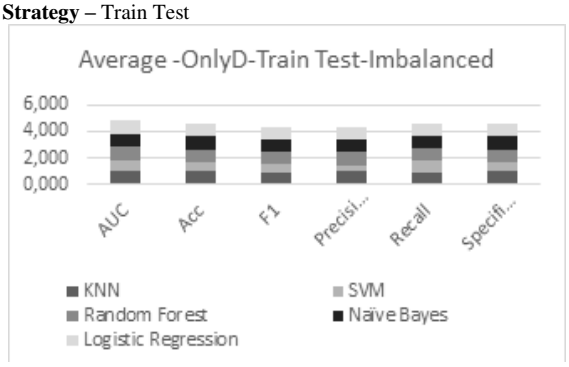


Figure 3. Imbalanced dataset experiment analysis only for diabetic records

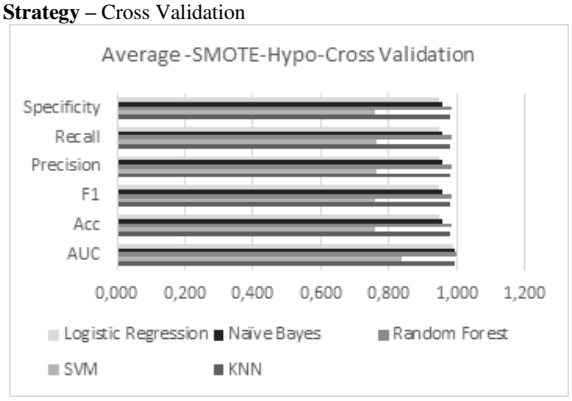
3.2. SMOTE Balanced Dataset experimentation

SMOTE oversampling method was used tobalance the dataset [28]-[30]. After SMOTE implementation the total number of records were 86.792 where 43396 were hypoglycemia detected records and 43.396 were no hypoglycemia detected records. Results were obtained as follows, SMOTE balanced dataset experiment analysis as shown in Figure 4.

3.3. ADASYN balanced dataset experimentation

ADASYN oversampling analysius was also done for getting balanced dataset [31]-[33]. After ADASYN implementation the total number of records were 86.704 where 43.308 were hypoglycemia detected records and 43.396 were no hypoglycemia detected records. Results were obtained as follows, ADASYN balanced dataset experiment analysis as shown in Figure 5.

Target – Hypoglycemia Detection
Dataset –SMOTE balanced dataset for Hypoglycemia imbalance
Strategy – Cross Validation



Target – Hypoglycemia Detection
Dataset –SMOTE balanced dataset for Hypoglycemia imbalance
Strategy – Train Test

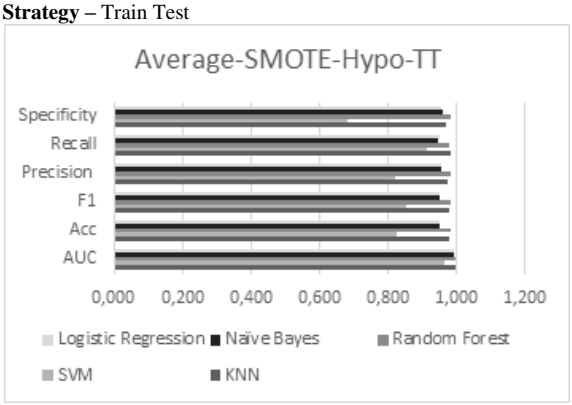
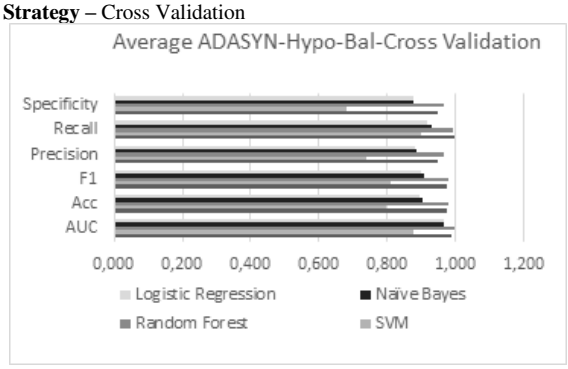


Figure 4. SMOTE balanced dataset experiment analysis

Target – Hypoglycemia Detection
Dataset – ADASYN balanced dataset for Hypoglycemia imbalance
Strategy – Cross Validation



Target – Hypoglycemia Detection
Dataset – ADASYN balanced dataset for Hypoglycemia imbalance
Strategy – Train Test

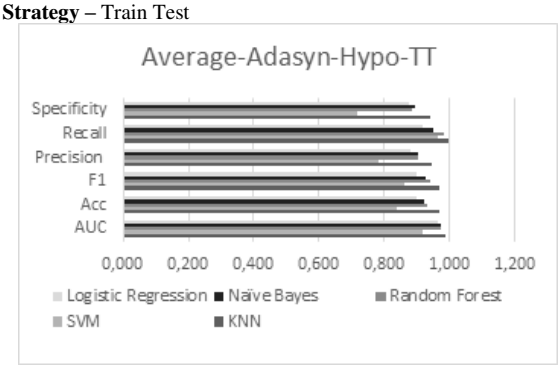


Figure 5. ADASYN balanced dataset experiment analysis

3.4. Comparative analysis

Three different considerations were done to come to conclusion. Comparative analysis of the metric values for different classifiers as shown in Figure 6.

Consideration 1: Average balanced and imbalanced based. The first consideration used avaraging on the scores obtained from different classifier metrics. All above observations conclude that the classifiers give comparatively similar metric values. The following graph shows the result of Average StackingC ensemble approach used. The average of all metric values is taken for all different experimentations done. Therefore we can see that the classifier based on its metric evaluation has following ratings. The Table 1 shows that the random forest classifier has highest metric values and therefore ranked 1.

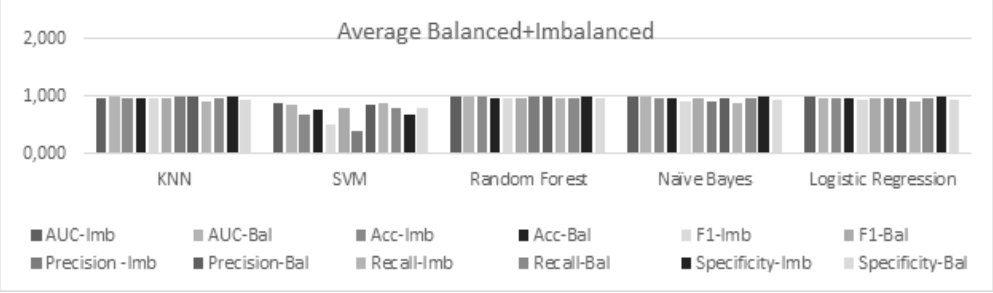


Figure 6. Comparative analysis of the metric values for different classifiers

Table 1. Classifier ratings according to metric evaluation consideration

Sr.no	Classifier	Relevance	Rank
1	Random Forest	Highest Metric values	1
2	KNN	Mid Metric values	2
3	Logistic Regression	Mid Metric values	3
4.	Naïve Bayes	Mid Metric values	4
5.	SVM	Lowest Metric values	5

Consideration 2: Confusion matrix based- the second consideration was taking into consideration false negative counts from confusion matrixIn healthcare anaysis along with true positive (TP) equal weightage is to be given to false negative (FN) count. The false negative count tells that though the hypoglycemia state is existing but it is not diagnosed, which is the state which may degrade the accuracy. The Table 2 shows the true positive, and false negative count for different classifier implementations with different strategies.

Table 2. Classifier confusion matrix true postive and false negative values obtained

Classifier	True Positive and False Negative counts	Imb-All Dataset	Imb-Only Diabetes Dataset	ADASYN Balanced Dataset	SMOTE Balanced Dataset
RF	TP	95.8	95.8	99.9	98.2
	FN	4.2	4.2	0.1	1.8
KNN	TP	93.5	93.5	99.6	97.5
	FN	6.5	6.5	0.4	2.5
SVM	TP	92.8	92.8	95.3	97.5
	FN	7.2	7.2	4.7	2.5
LR	TP	91.8	91.8	89.5	94.9
	FN	8.2	8.2	10.5	5.1
NB	TP	90.2	90.2	81.8	94
	FN	9.8	9.8	19.2	6

Table 3 shows the ranking based on the highest true positive and lowest false negative count. Random forest classifier is at the rank 1 and Naïve Bayes at the last. Random forest being iterative decision tree ensembling give more true poistive counts.

Table 3. Classifier ratings according to confusion matrix true postive and false negative consideration

Sr.no	Classifier	Relevance	Rank
1	Random Forest	Highest Preference	1
2	KNN	Mid Preference	2
3	Logistic Regression	Mid Preference	3
4.	SVM	Mid Preference	4
5.	Naïve Bayes	Lowest Preference	5

Consideration 3: Train and test time based-the third consideration was taking time into consideration.
The average train time and test time is calculated based on all the experimentations done for following categories,

1. All imbalanced dataset implementations
2. ADASYN balanced dataset
3. SMOTE balanced dataset

Table 4 gives the glimpse of train-test timing required for different classifier implementations. Whenever any implementations are to be done the timing for execution is the factor which cannot be neglected. The Table 4 shows Naïve Bayes having least time of execution and SVM having the maxmum time.

Table 4. Classifier ratings according to time consideration

	ADASYN		SMOTE		Imbalanced	
Classifier	Average Train Time	Average Test Time	Average Train Time	Average Test Time	Average Train Time	Average Test Time
KNN	3.396	11.439	4.424875	11.99013	2.532	8.678
SVM	25.877	0.856	33.4774375	0.877875	19.527	0.679
Random Forest	5.591	0.459	5.0714375	0.404625	4.130	0.321
Naïve Bayes	0.562	0.077	0.6616875	0.083125	0.437	0.059
Logistic Regression	7.591	0.050	8.733625	0.062125	6.749	0.044

3.5. Discussions

The above implementation and results show that only classifier-based evaluation is not sufficient for any conclusion. Especially when analysis is to be done in healthcare sector then ensemble approach plays very important role. The different metric scores with false negative counts give very clear idea that which classifier suits well for the analysis to be done. Also, time consideration can be done if is acute critical diagnosis in healthcare. An Ensemble approach is the way to get more accurate results.

3.6. Comparison with the existing implementations

The implementation done by authors in paper [25] only focuses on different classifier cumulative scores, while the proposed ensemble approach takes into consideration different machine learning classifiers, evaluation methods like cross fold, stratified cross fold, train test and repeat train test. This multi-step ensemble gives more accurate results. Also, in the articles [24]-[27] only use of different classifier and performance metrics is done, while in the proposed method not only the performance metric score but two more considerations are focused on viz. false negative counts and execution time. The stacking C approach to get final metric scores, confusion matrix for false negative counts and time considerations makes the results more concrete and clearer for analysis.

4. CONCLUSION

Machine learning implementation for 5 different classifier viz. random forest, KNN, logistic regression, Naïve Bayes and SVM was thus done using innovative ensemble method of machine learning. SMOTE and ADASYN oversampling algorithms was used for balancing the imbalance due to hypoglycemia detected attribute. The results were concluded considering 3 important points viz. experimentation metric analysis, train test time comparison and TP and FN values. The 3-stage ensemble experimentation concluded that random forest, KNN, logistic regression and Naïve Bayes were good to be considered for hypoglycemia detection. Random forest classifier found to be the most stable classifier for all strategies. Also, we can conclude that FN count should be taken into consideration as disease not detected correctly may lead to serious conditions. Considering FN count it was also found that the SMOTE oversampling method gives more accuracy in terms of less FN counts. An ensemble approach gives a better understanding of the implementation done which further helps in proper decision making too. So rather than only comparing the classifier values obtained for different performance metrics, it is always better to use appropriate ensemble approach best suited for application. In healthcare data analytics major concern is imbalanced dataset which can be an important consideration for an ensemble approach for healthcare data analytics.

REFERENCES

[1] P. E. Cryer, *Hypoglycemia in Diabetes*. Textbook of Diabetes, pp. 528-545, 2010, doi: 10.1002/9781444324808.ch33.

[2] D. Sierra-Sosa, B. Garcia-Zapirain, C. Castillo, I. Oleagordia, R. Nuno-Solinis, M. Urtaran-Laresgoiti, and A. Elmaghraby, "Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs," *IEEE Transactions on Industrial Informatics,* vol. 15, no. 10, pp. 5682–5689, 2019, doi: 10.1109/tii.2019.2919168.

[3] M. Porumb, S. Stranges, A. Pescapè, and L. Pecchia "Precision medicine and artificial intelligence: A Pilot study on deep learning for hypoglycemic events detection based on ECG," *Sci Rep*, vol. 10, no. 170, 2020, doi: 10.1038/s41598-019-56927-5.

[4] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," in *IEEE Access*, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[5] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[6] S. Ray, "A quick review of machine learning algorithms," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.

[7] R. Ghorbani, R. Ghousi, A. Makui, and A. Atashi, "A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset," *IEEE Access*, vol. 8, pp. 141066-141079, 2020, doi: 10.1109/access.2020.3013320.

[8] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," *in IEEE Access*, vol. 8, pp. 171263-171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[9] E. R. Fernandes, A. C. P. L. F. De Carvalho, and X. Yao, "Ensemble of classifiers based on multiobjective genetic sampling for imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1104-1115, 2020, doi: 10.1109/tkde.2019.2898861.

[10] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based idss on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150-32162, 2020, doi: 10.1109/access.2020.2973219.

[11] U. R. Salunkhe, and S. N. Mali, "Classifier ensemble design for imbalanced data classification: a hybrid approach," *Procedia Computer Science*, vol. 85, pp. 725-732, 2016, doi: 10.1016/j.procs.2016.05.259.

[12] T. Hasanin, T. M. Khoshgoftaar, and J. L. Leevy, "A comparison of performance metrics with severely imbalanced network security big data," 2019 IEEE 20th *International Conference on Information Reuse and Integration for Data Science (IRI),* Los Angeles, CA, USA, 2019, pp. 83-88, doi: 10.1109/IRI.2019.00026.

[13] M. Naghshvarianjahromi, S. Kumar, and M. J. Deen, "Brain-inspired intelligence for real-time health situation understanding in smart e-health home applications," *in IEEE Access*, vol. 7, pp. 180106-180126, 2019, doi: 10.1109/ACCESS.2019.2958827.

[14] K. Anam, H. Ismail, F. S. Hanggara, C. Avian, and S. B. Worsito, "Cross validation configuration on K-NN for finger movements using EMG signals," 2021 *International Conference on Instrumentation, Control, and Automation (ICA),* 2021, pp. 17-21, doi: 10.1109/ICA52848.2021.9625699.
[15] D. Javale, S. Desai, "Dataset for people for their blood glucose level with their superficial body feature readings," *IEEE Dataport*, July 7, 2021, doi: 10.21227/c4pp-6347.
[16] A. H. Syed and T. Khan, "Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: a retrospective cross-sectional study," in *Access*, vol. 8, pp. 199539-199561, 2020, doi: 10.1109/ACCESS.2020.3035026.
[17] M. A. Awal *et al*., "An early detection of asthma using BOMLA detector," in *IEEE Access*, vol. 9, pp. 58403-58420, 2021, doi: 10.1109/ACCESS.2021.3073086.
[18] J. R. Campos, E. Costa, and M. Vieira, "Improving failure prediction by ensembling the decisions of machine learning models: a case study," in *IEEE Access*, vol. 7, pp. 177661-177674, 2019, doi: 10.1109/ACCESS.2019.2958480.
[19] K. Zarkogianni, M. Athanasiou, A. C. Thanopoulou, and K. S. Nikita, "Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication," in *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1637-1647, Sept. 2018, doi: 10.1109/JBHI.2017.2765639.
[20] A. Sharma and H. Wehrheim, "Testing machine learning algorithms for balanced data usage," *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*, Xi'an, China, 2019, pp. 125-135, doi: 10.1109/ICST.2019.00022.
[21] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation metrics for evaluating classification performance on imbalanced data," *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Tangerang, Indonesia, 2019, pp. 14-18, doi: 10.1109/IC3INA48034.2019.8949568.
[22] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [Research Frontier]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 59-76, 2018, doi: 10.1109/mci.2018.2866730.
[23] M.-P. Hosseini, A. Hosseini, and K. Ahi, "A review on machine learning for EEG signal processing in bioengineering," in *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 204-218, 2021, doi: 10.1109/RBME.2020.2969915.
[24] P. Sun, D. Wang, V. C. Mok, and L. Shi, "Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading," in *IEEE Access*, vol. 7, pp. 102010-102020, 2019, doi: 10.1109/ACCESS.2019.2928975.
[25] M. Gramajo, L. Ballejos, and M. Ale, "Seizing requirements engineering issues through supervised learning techniques," in *IEEE Latin America Transactions*, vol. 18, no. 07, pp. 1164-1184, July 2020, doi: 10.1109/TLA.2020.9099757.
[26] J. K. Agor, N. L. P. S. P. Paramita, and O. Y. Özaltın, "Prediction of sepsis related mortality: an optimization approach," in *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 11, pp. 4207-4216, Nov. 2021, doi: 10.1109/JBHI.2021.3096470.
[27] T. Rahman *et al*., "Development and validation of an early scoring system for prediction of disease severity in COVID-19 using complete blood count parameters," in *IEEE Access*, vol. 9, pp. 120422-120441, 2021, doi: 10.1109/ACCESS.2021.3105321.
[28] J. Wei, Z. Lu, K. Qiu, P. Li, and H. Sun, "Predicting drug risk level from adverse drug reactions using SMOTE and machine learning approaches," in *IEEE Access*, vol. 8, pp. 185761-185775, 2020, doi: 10.1109/ACCESS.2020.3029446.
[29] K. Kim, "Noise avoidance SMOTE in ensemble learning for imbalanced data," in *IEEE Access*, vol. 9, pp. 143250-143265, 2021, doi: 10.1109/ACCESS.2021.3120738.
[30] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," in *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3064084.
[31] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Adaptive synthetic-nominal (ADASYN-N) and adaptive synthetic-KNN (ADASYN-KNN) for multiclass imbalance learning on laboratory test data," *2018 4th International Conference on Science and Technology (ICST)*, 2018, pp. 1-6, doi: 10.1109/ICSTC.2018.8528679.
[32] C. Lu, S. Lin, X. Liu, and H. Shi, "Telecom fraud identification based on ADASYN and random forest," *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, 2020, pp. 447-452, doi: 10.1109/ICCCS49078.2020.9118521.
[33] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

BIOGRAPHIES OF AUTHORS

Deepali Pankaj Javale has completed her M. Tech in information technology and currently pursuing her PhD in Computer Engineering from MIT World Peace University, Pune, India. She is working as Assistant Professor in School of Computer Engineering and Technology, MIT World Peace University, Pune, India since 2008. Having more than 23 years of experience in teaching, she has more than 20 publications on her account with good number of citations. Her area of interest are embedded system, IoT, artificial intelligence, machine learning, data analytics and healthcare analytics. She can be contacted at: deepali.javale@mitwpu.edu.in.

Dr. Sharmishta Suhas Desai has received Ph.D in Computer Engineering from the University of Pune in 2017. She is working as an Associate Professor in the School of Computer Engineering and Technology, MIT World Peace University, Pune since 2008. Her main research work focuses on Big Data, Machine Learning, Artificial Intelligence, Natural Language Processing, Data Analytics and Data Mining. She has more than 18 years of teaching experience. She has received different research grants and consultancy on her name. She can be contacted at: sharmishta.desai@mitwpu.edu.in.