Healthcare Critical Diagnosis Accuracy: A Proposed Machine Learning Evaluation Metric for Critical Healthcare Analysis

Original Scientific Paper

Deepali Pankaj Javale
School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India
deepali.javale@mitwpu.edu.in

Sharmishta Desai
School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India
sharmishta.desai@mitwpu.edu.in

Abstract *– Since at least a decade, Machine Learning has attracted the interest of researchers. Among the topics of discussion is the application of Machine Learning (ML) and Deep Learning (DL) to the healthcare industry. Several implementations are performed on the medical dataset to verify its precision. The four main players, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), play a crucial role in determining the classifier's performance. Various metrics are provided based on the main players. Selecting the appropriate performance metric is a crucial step. In addition to TP and TN, FN should be given greater weight when a healthcare dataset is evaluated for disease diagnosis or detection. Thus, a suitable performance metric must be considered. In this paper, a novel machine learning metric referred to as Healthcare-Critical-Diagnostic-Accuracy (HCDA) is proposed and compared to the well-known metrics accuracy and ROC_AUC score. The machine learning classifiers Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB) are implemented on four distinct datasets. The obtained results indicate that the proposed HCDA metric is more sensitive to FN counts. The results show, that even if there is rise in %FN for dataset 1 to 10.31 % then too accuracy is 83% ad HCDA shows correlated drop to 72.70 %. Similarly, in dataset 2 if %FN rises to 14.80 for LR classifier, accuracy is 78.2 % and HCDA is 63.45 %. Similar kind of results are obtained for dataset 3 and 4 too. More FN counts result in a lower HCDA score, and vice versa. In common exiting metrics such as Accuracy and ROC_AUC score, even as the FN count increases, the score increases, which is misleading. As a result, it can be concluded that the proposed HCDA is a more robust and accurate metric for Critical Healthcare Analysis, as FN conditions for disease diagnosis and detection are taken into account more than TP and TN.*

Keywords*: Machine Learning, Performance Metrics, Accuracy, ROC_AUC, Healthcare, True Positive, True Negative, False Negative*

1. INTRODUCTION

Machine learning has proved beneficial in a variety of fields. The analysis of healthcare data is also gathering popularity. But it also has its difficulties [1]. Different Ensemble methods [2] have also proven superior for achieving high precision. When discussing Machine Learning, performance evaluation metrics play an important role in determining how closely the parameters influence the target field. In supervised learning, we specify a target field, train the model with various classifiers, and test it on a small sample of records using the same parameters. The performance metrics accuracy, F1-Score, Precision, Recall, and ROC_AUC Score play a crucial role in classifier implementation comparisons based on performance metrics. When discussing these metrics, the perplexity matrix is used to calculate their scores. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are the four key actors from the confusion matrix that aid in calculating the metric scores. Accuracy [2] and ROC_AUC score metrics are typically used in healthcare data analytics [3-6]. When discussing these metrics, it is observed that false negative counts are not taken into account, which should be one of the most important considerations when dealing with essential healthcare analytics. The ROC_AUC score takes into account the number of false negatives and is therefore superior. In addition, the Recall and F1-Score metrics can be considered because they include FN as one of their key decision-making parameters.

When all of the aforementioned metrics are considered, it is observed that although the number of FNs increases,

the metric value also increases, which is contradictory. The number of false negatives in a critical healthcare analysis should be minimal. For instance, if a heart attack prediction is made and the false negative readings indicate that even though it is a heart attack state, it was not correctly predicted, this is extremely dangerous for the patient. The false positive state is tolerable because it may be a false alarm for a critical situation, but it may not pose a life-threatening threat. Taking all of these factors into account.

Further, the study was to explore more performance metrics in machine learning. Just relying on an accuracy score was not a good choice. In one of the articles, the authors have given a comparative study of different metrics used in machine learning for imbalanced datasets. 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 effect. [7]

Different ensemble approaches for machine learning viz. bagging, Breiman boosting, and Freund boosting. Imbalanced datasets are mainly to be taken into consideration. Different metrics for imbalanced datasets were discussed and experimented with. AUC was considered to be the most robust [8]. When the results of various implementations for different metric values such as accuracy, ROC_AUC score, F1 score, Precision score, and Recall score were compared, it was discovered that the percentage of False Negatives was increasing while the accuracy was increasing. Consequently, the Health-Critical-Diagnosis-Accuracy (HCDA) metric was conceived. Using various classifiers of Machine Learning, four distinct healthcare datasets were implemented. Comparing the implementation results for metrics accuracy, ROC_AUC Score, and the HCDA state-of-the-art metric revealed that HCDA produced more accurate results.

The paper includes below given contributions,

- Four different datasets are used for disease or medical condition detection for which the statistics are given in section 2.1.
- Different Machine Learning performance evaluation metrics are discussed in section 2.2 which are further used for results and conclusions.
- A state-of-art metric is proposed, Healthcare-Critical-Diagnosis-Accuracy(HCDA) which is given in section 2.3
- Comparative analysis of different machine metrics from section 2.2 and HCDA metric from 2.3 are compared together.
- In section 3 results for all experimentation are given and discussion on it is done.
- Lastly, a conclusion is stated which shows the significance of the proposed work.

Need for Proposed Work :

Despite the various performance evaluation metrics provided by the machine learning community, it has been observed that critical areas, such as Critical Healthcare Analysis, require additional attention and development. In numerous instances, an ensemble approach utilising machine learning proves to be beneficial. In critical healthcare analysis, emphasis must be placed not only on True positive and True negative cases, but also on False negative cases. False negative contricution must be understood when calculating the metric value. Therefore, the proposed work is an effort to focus more on False Negative counts in order to achieve greater accuracy in Healthcare Analytics, thereby reducing the risk of death in critical conditions such as stroke and heart attack.

2. METHOD

2.1. DATASETS USED

Machine learning implementation was done on 4 different datasets. The first 2 datasets used were the diabetes dataset while the 3rd dataset was the stroke prediction dataset and the 4th was the Heart Failure Clinical Record dataset.

The statistics and description of the four different datasets used are given in Table 1.

Table 1. Summary of Datasets Used

Dataset Number	Title	Attributes	Fields considered	Number of Records	Description
1.	Dataset for People For Their Blood Glucose Level With Their Superficial Body Feature Readings [9]	10	Diastolic BP,Systolic BP, Heartrate, Shivering BodyTemperature, Hypoglycemia(Target Field)	70000	The given Dataset is a record of different age groups of people either diabetic or non-diabetic for their blood glucose level reading with superficial body features like body temperature, heart rate, blood pressure, etc.
2.	Diabetes Data Set [10]	9	Pregnancies,BloodGlucose blood pressure, SkinThickness, Outcome (Target Field)	2000	Predict a Model to detect Whether Person has Diabetes or Not
3.	STROKE PREDICTION DATASET [11]	12	Id, Gender, Age, Hypertension, Heart disease, Ever marrried, Worktype Stroke(Target Field)	5111	11 clinical features for predicting stroke events
4.	HEART FAILURE PREDICTION [12]	13	Age, anaemia, high blood pressure, creatinine, phosphokinase, diabetes, ejection fraction, platelets, sex, serum creatinine, serum sodium, smoking, [target] death event(Target Field)	300	12 clinical features for predicting death events

Taking into consideration the comparative analysis from Table 2 for supervised machine learning classifiers, four machine learning classifiers viz. Random Forest (RF), Support Vector Machine(SVM), Naïve Bayes (NB) and Logistic regression(LR) [13-17] were used for supervised classification using Machine Learning [2, 18,19] on the above datasets. The train test method with Stratified Crossfold with k=10 strategy was used for classifier experimentation. The chances of missing any of the train or test records are eliminated in Stratified Crossfold [20-24] mechanism of Machine Learning. The commonly used machine learning evaluation metrics based on the confusion matrix are Accuracy, F1-Score, Precision, Recall, and ROC_AUC. Here according to confusion matrix TP means True Poistive which means correct prediction for true/positive values. TN means True predictions for False/Negative values. FP means wrong predictions for True/Positive values. FN means wrong predictions for False/Negative Values

$$Accuracy = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+TN+FP+FN})} \quad (1)$$

$$Precision = \frac{TP}{(\mathrm{TP+FP})} \quad (2)$$

$$Recall = \frac{TP}{(\mathrm{TP+FN})} \quad (3)$$

$$F1 - Score = \frac{(2(\mathrm{P*Q}))}{(\mathrm{P+Q})} \quad (4)$$

$$TPR = \frac{TP}{(\mathrm{TP+FN})} \quad (5)$$

$$FPR = \frac{FP}{(\mathrm{FP+TN})} \quad (6)$$

$$ROC\_AUC = (Eq\ (5))/(Eq\ (6))$$

The evaluation metrics used for comparative analysis were Accuracy, F1-Score, Precision, Recall, and ROC_AUC from Machine Learning. Along with these metrics, the proposed HCDA metric is also used. The Accuracy, F1-Score, Precision, Recall, and ROC_AUC are then compared with HCDA a state-of-art metric used.

$$HCDA = \frac{(\mathrm{TP+TN*100}}{(\mathrm{TP+TN+FP+FN})} - \frac{(\mathrm{FP*100})}{(\mathrm{TP+TN+FP+FN})} \quad (7)$$

Based on the confusion matrix parameters TP, TN, FP and FN consideration for the value calculation of respective metrics, the metrics and their mapping [25-27] are shown in Table 3.

Table 3. Machine Learning and HCDA metric mapping with the confusion matrix key players

Metric	True Positive (*TP*)	True Negative (*TN*)	False Positive (*FP*)	False Negative (*FN*)
Accuracy	√	√	√	√
Precision	√	-	√	-
Recall	√	-	-	√
F1-Score	√	-	√	√
ROC_AUC	√	√	√	√
HCDA	√	√	√	√

The above table shows the use of TP, TN, FP, or FN for the metric value calculation. In further discussions, False Negative (FN) is considered to be an important player as in critical healthcare analysis the condition of false negative is considered to be more alarming. If the critical health state is taking place and it's not indicated then such a situation is termed false negative which should not be tolerated.

If we consider a dataset for stroke diagnosis, then

TP- The patient is undergoing stroke and is correctly diagnosed

TN- The patient is not undergoing a stroke and is correctly diagnosed.

FP- The patient is wrongly diagnosed as undergoing a stroke.

FN – The patient is wrongly diagnosed as not undergoing a stroke.

If we look at the above case study of healthcare critical analysis, then it is observed that TP and TN are correct to be found but along with it, the most important is the FN count. If the FN count goes high it means the system is failing in classifying critical health conditions. Compared to it if FP count goes high then too it may not be a risk to the patient.

Thus, from Table 3, it is clear that for healthcare analysis metrics like Precision and Recall should not be considered for Critical Healthcare Analysis. The methodology implementations are done on the datasets given in Table 1 and we continued with metrics Accuracy, Precision, Recall, F1-Score, ROC_AUC [28,29] and HCDA proving that how HCDA is a better metric for critical healthcare analysis as compared to all other metrics. The significance of TP, TN and FN is to be justified using a mathematical model or graph comparison.

2.3. PROPOSED METRIC

The state-of-art metric is proposed named HCDA (Healthcare-Critical-Diagnosis-Accuracy). The HCDA is the percentage difference between the sum of true positives and true negatives and the percentage of false negatives. In critical healthcare diagnosis along with true positive and true negative more importance should be given to false negative. The false negative state mentions that even the critical state is occurring then diagnosis is not done which is considered to be more dangerous. The false positive count can be neglected as false alarms can be tolerated. When we use other Machine Learning Metrics then it is observed that we need to get the values of various metrics like precision, accuracy, F1-Score, Recall, and ROC_AUC score. Either of the strategies have to be used to come to conclude like stacking-C, Max vote, or Average. If we use HCDA the only metric gives accurate results.

Where,

$HCDA$ - Healthcare - Critical - Diagnosis - Accuracy (Proposed metric value)

TP - True Positive (Correct Diagnosis counts. Critical State occurring with proper indication)

TN - True Negative(Correct Diagnosis Count. Critical State occurring with no indication)

FN – False Negative (Wrong Diagnosis. Critical State occurring but no indication)

The HCDA metric focuses on true positive and true negative for calculating accuracy and in the same way removes the percentage of false negative counts to get more accurate results.

The steps for calculating HCDA metric score are given in Algorithm 1.

Algorithm 1 :

Result: HCDA Metric score

List: Support Vector Machine, Logistic Regression, Naïve Bayes and Random Forest

Dataset: 4 Datasets from Table 1

Steps:

for <for each algorithm in **List**> ***do***

for <for each dataset from **Dataset**> ***do***

Implement Machine Learning Algorithm

Get Confusion Matrix Parameter values for TP, TN, FP, and FN

Calculate HCDA by the given formula

$$HCDA = \frac{TP + TN * 100}{TP + TN + FP + FN} - \frac{FP * 100}{TP + TN + FP + FN}$$

end

end

2.4. COMPARATIVE ANALYSIS

Comparative analysis was done between the evaluation metric values obtained for Machine Learning metrics used and HCDA, the state-of-art metric proposed.

The comparative analysis proposes the effectiveness of the HCDA metric

The workflow architecture for the proposed work and experimentation is shown in Fig 1.

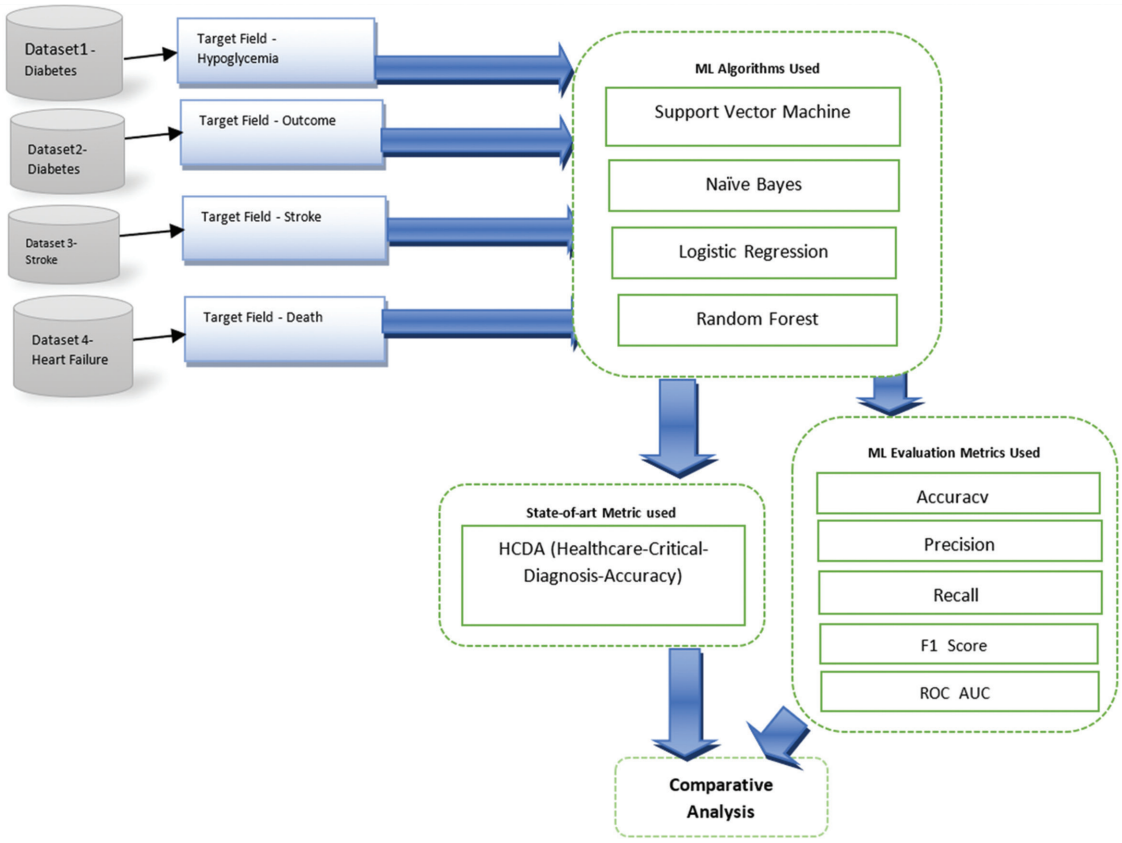

Fig 1. The Workflow Architecture of the Implementation Done

3. RESULTS AND DISCUSSION

The results obtained were the comparative chart for the values of different metric values after the implementation of classifier execution on the four different healthcare datasets used. The datasets were picked up from the Kaggle repository and IEEE data port.

Table 4 shows the true positive, true negative, false negative, and false positive counts obtained after classifier execution on the respective dataset.

Based on the values of TP, TN, FP, and FN the metric values for ML classifier execution were calculated for different machine learning metrics and the HCDA metric proposed. All the metric values are considered to be in percentage.

Table 5 shows the %(TP+TN) and %FN values along with the value generated for the HCDA metric and the standard machine learning metrics viz. accuracy, precision, recall, F1-Score, and ROC_AUC score. Though the accuracy score gives a good score, it can be seen that the accuracy score is directly correlating with %(TP+TN) only. The FN value rise does not affect the accuracy score, while HCDA shows variations accordingly.

Table 4. The TP, TN, FN, and FP values for different machine learning classifier implementations

Dataset / Metric	Classifier	TN	TP	FP	FN
D1 (Base Dataset)	RF	61869	8673	19	382
	SVM	57157	1738	4731	7317
	LR	61603	8259	285	796
	NB	60912	8091	976	964
D2 - Diabetes	RF	1303	642	13	42
	SVM	708	363	608	321
	LR	1177	388	139	296
	NB	1032	468	284	216
D3 - Stroke	RF	4841	5	20	244
	SVM	4534	21	327	228
	LR	4852	2	9	247
	NB	4673	46	188	203
D4 - Heart Failure	RF	179	66	24	30
	SVM	185	52	18	44
	LR	184	61	19	35
	NB	182	69	21	27

Table 5. Different metric values obtained for Machine Learning classifier implementations

Dataset/ Metric	Classifier	%(Tp+TN)	%FN	HCDA	Accuracy	Precision	Recall	F1score	ROC
D1 (Base Dataset)	RF	99.43	0.54	98.90	99.4	99.4	99.4	99.4	99.4
	SVM	83.02	10.31	72.70	83	80.8	83	81.8	61.6
	LR	98.48	1.12	97.35	98.5	98.5	98.5	98.5	99.6
	NB	97.27	1.36	95.91	97.3	97.3	97.3	97.3	99.5
D2 - Diabetes	RF	97.25	2.10	95.15	97.3	97.3	97.3	97.2	99
	SVM	53.55	16.05	37.50	53.5	58.1	53.5	54.7	56.5
	LR	78.25	14.80	63.45	78.2	77.8	78.2	77.5	83.3
	NB	75.00	10.80	64.20	75	75.7	75	75.3	82.3
D3 - Stroke	RF	94.83	4.77	90.06	94.8	91.5	94.8	92.8	74.5
	SVM	89.14	4.46	84.68	89.1	90.9	89.1	90	56.8
	LR	94.99	4.83	90.16	95	91.4	95	92.8	65.5
	NB	92.35	3.97	88.38	92.3	92.1	92.3	92.2	80.6
D4 - Heart Failure	RF	81.94	10.03	71.91	81.9	81.7	81.9	81.8	88
	SVM	79.26	14.72	64.55	79.3	78.7	79.3	78.3	84
	LR	81.94	11.71	70.23	81.9	81.5	81.9	81.5	85.5
	NB	83.95	9.03	74.92	83.9	83.7	83.9	83.8	90.8

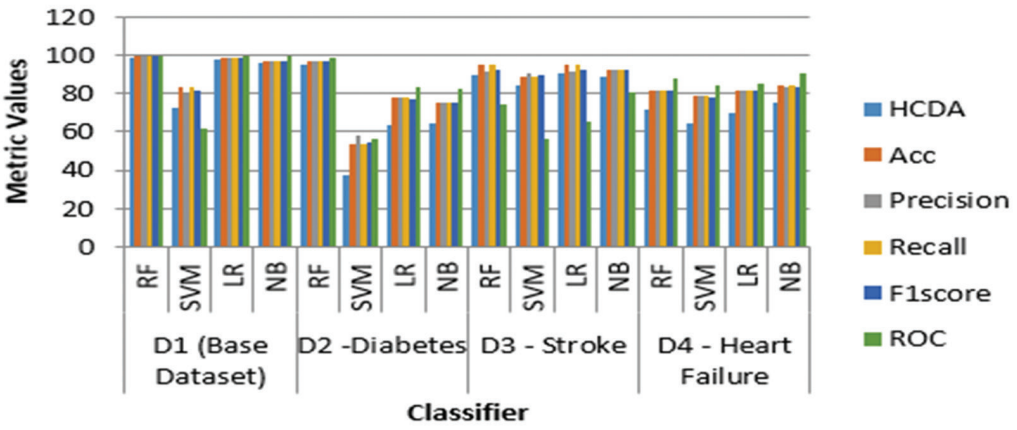

Fig. 2. Comparison of HCDA metric with Machine Learning Metric values.

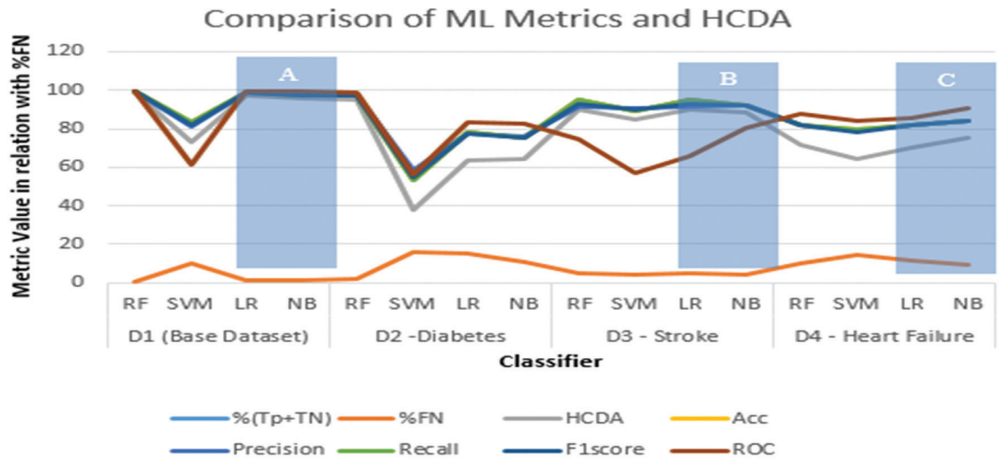

Fig 3. Comparison of HCDA metric with Machine Learning Metric values.

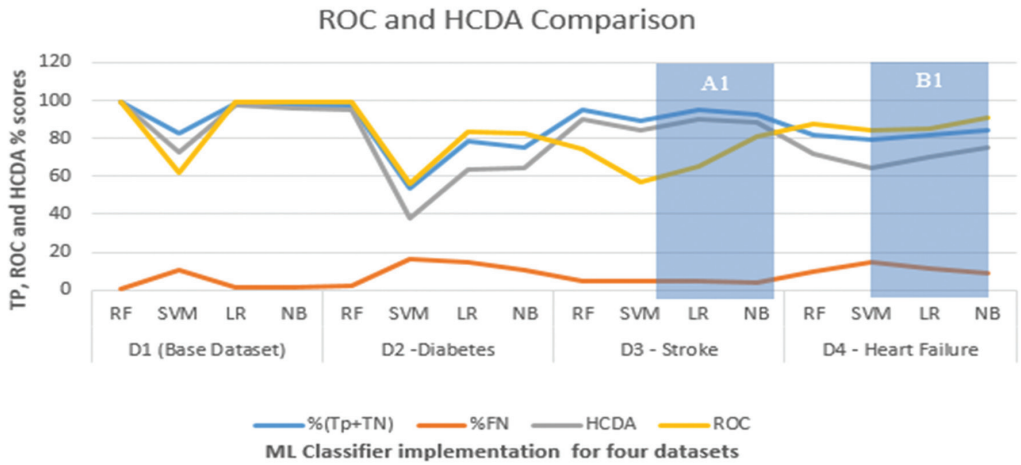

Fig 4. Comparative Analysis for ROC and HCDA metric

Discussion 1 : Fig. 2 compares various metrics, including precision, accuracy, F1-score, Recall, and ROC score, to the proposed HCDA metric. In healthcare diagnosis, precision, accuracy, F1-score, Recall, and ROC score metrics are likely adopted to reach a conclusion [4,5]. The HCDA value differs from the machine learning metrics' values. The HCDA metric value is relatively low compared to other metric values. The ROC is regarded as the most reliable metric for healthcare analysis. However, if we observe the ROC and HCDA scores attentively, we will notice that in many instances the HCDA score is greater than the ROC score. The primary topic of discussion is not achieving a higher accuracy score, but rather achieving the most accurate score in relation to the percentages of True Positives and True Negatives as well as False Negatives.

Observing Fig. 3 closely reveals that the HCDA score is highly correlated with $\%FN$ values. The greater the value of $\%FN$, the lower the HCDA score, whereas there is no correlation between $\%FN$ value and ROC score. The shaded area A in Fig. 3 indicates that the $\%FN$ count is greater. Under this condition, the ROC score and HCDA score both decrease significantly. In addition, the shaded portion B reveals that the $\%FN$ score is significantly lower than the ROC score, while the HCDA score has a much stronger correlation with the $\%FN$ count. The shaded section C indicates an increase in $\%FN$. In such a scenario, the accuracy score should decrease, but the ROC score is high and the HCDA score remains stable in correlation with $\%FN$.

Here the $\%FN$ score means,

For Dataset 1 – A hypoglycemia state occurs but is not detected

For Dataset 2 - Diabetes positive but not detected

For Dataset 3 - Stroke occurs but is not detected

For Dataset 4 – A heart failure state occurs but is not detected

From the above four respective FN states for 4 different datasets, it is clear that the FN count should be the most important parameter along with TP and TN.

Discussion 2: ROC is typically regarded as the most reliable metric, and healthcare is no exception. Figure 4 depicts a comparison of the ROC score and HCDA score in relation to the percentage of true positive, true negative, and false negative scores. It demonstrates that the %(true positive + true negative) score and %false negative score, despite being identical in many instances,

reflect distinct ROC scores, whereas HCDA indicates a very close correlation between them.

Observing Fig. 4, part A1 reveals that the percentage of FN is relatively low compared to the percentage of $TP+TN$. In this scenario, a high accuracy score should be reflected by the HCDA score and not the ROC score. In addition, the shaded portion B1 displays an increase in $\%FN$ and a minor decrease in $\%(TP+TN)$, which should result in a decrease in the accuracy score. However, this is best reflected by the HCDA score and not the ROC score. Fig. 4 demonstrates that the HCDA metric is consistent with respect to the %false negative score. For Acute/Critical healthcare, detection $\%FN$ is a very essential factor, and HCDA is a self-sufficient metric for determining the classifier's performance.

4. CONCLUSION

Consequently, the implemented research demonstrates that the proposed state-of-the-art metric HCDA is more robust and superior for critical healthcare analysis than other machine Learning metrics such as Accuracy, Precision, Recall, F1-Score, and ROC. The execution of various classifiers, namely Random Forest, Nave Bayes, Support Vector Machine, and Logistic Regression, leads to the conclusion that the HCDA metric is more accurate for critical healthcare diagnosis. The four datasets employed were the datasets for critical healthcare analysis in which acute state detection is the primary objective. All experiments demonstrated that the proposed HCDA metric is the only self-sufficient metric capable of producing accurate classification decisions. For the proposed metric HCDA, the accuracy will increase if the number of false negatives decreases. The HCDA demonstrates a very strong correlation with the True Positive, True Negative, and False Negative values, which is essential for conducting critical healthcare analyses. The minority class, which is represented by the false negative count, should therefore be weighed equally with the true positive and true negative tallies. If the HCDA metric is used for decision-making in critical healthcare analysis, such as heart failure or stroke, then putting more emphasis on false negative cases will prevent or reduce the occurrence of severe conditions that are not detected. This demonstrates that the HCDA metric has the potential to revolutionise acute state detection analysis in healthcare. Despite the fact that the scope of this study is limited to Critical Healthcare Analysis, the HCDA metric can be applied in sectors such as the aerospace and military that place a premium on false-negative conditions.

5. REFERENCES

[1] A. Qayyum, J. Qadir, M. Bilal, A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey", IEEE Reviews in Biomedical Engineering, Vol. 14, 2021, pp. 156-180.

[2] M.-P. Hosseini, A. Hosseini, K. Ahi, "A Review on Machine Learning for EEG Signal Processing in Bioengineering", IEEE Reviews in Biomedical Engineering, Vol. 14, 2021, pp. 204-218.

[3] N. Y. Philip, M. Razaak, J. Chang, S. M. M. O'Kane, B. K. Pierscionek, "A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes", IEEE Access, Vol. 10, 2022, pp. 13460-13471.

[4] M. Habib, Z. Wang, S. Qiu, H. Zhao, A. S. Murthy, "Machine Learning Based Healthcare System for Investigating the Association Between Depression and Quality of Life", IEEE Journal of Biomedical and Health Informatics, Vol. 26, No. 5, 2022, pp. 2008-2019.

[5] N. Reamaroon, M. W. Sjoding, K. Lin, T. J. Iwashyna, K. Najarian, "Accounting for Label Uncertainty in Machine Learning for Detection of Acute Respiratory Distress Syndrome", IEEE Journal of Biomedical and Health Informatics, Vol. 23, No. 1, 2019, pp. 407-415.

[6] Y. Dong et al. "A Polarization-Imaging-Based Machine Learning Framework for Quantitative Pathological Diagnosis of Cervical Precancerous Lesions", IEEE Transactions on Medical Imaging, Vol. 40, No. 12, 2021, pp. 3728-3738.

[7] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, "A Comparison of Performance Metrics with Severely Imbalanced Network Security Big Data", Proceedings of the IEEE 20th International Conference on Information Reuse and Integration for Data Science, Los Angeles, CA, USA, 2019, pp. 83-88.

[8] M. Naghshvarianjahromi, S. Kumar, M. J. Deen, "Brain-Inspired Intelligence for Real-Time Health Situation Understanding in Smart e-Health Home Applications", IEEE Access, Vol. 7, 2019, pp. 180106-180126.

[9] D. Javale, S. Desai, "Dataset for People for their Blood Glucose Level with their Superficial body feature readings", IEEE Dataport, 2021.

[10] "Diabetes Data Set (Version 1) [Predict a Model to detect Person has Diabetes or Not]", https://www.kaggle.com/datasets/vikasukani/diabetes-dataset (accessed: 2023)

[11] "Stroke Prediction Dataset (Version 1) [11 clinical features for predicting stroke events]", https://

www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed: 2023)

[12] D. Chicco, G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone", BMC Medical Informatics and Decision Making, Vol. 20, No. 16, 2020.

[13] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", IEEE Access, Vol. 8, 2020, pp. 107562-107582.

[14] J. Faouzi et al. "Machine Learning-Based Prediction of Impulse Control Disorders in Parkinson's Disease From Clinical and Genetic Data", IEEE Open Journal of Engineering in Medicine and Biology, Vol. 3, 2022, pp. 96-107.

[15] S. A. -F. Sayed, A. M. Elkorany, S. Sayed Mohammad, "Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity", IEEE Access, Vol. 9, 2021, pp. 135697-135707.

[16] J. R. Campos, E. Costa, M. Vieira, "Improving Failure Prediction by Ensembling the Decisions of Machine Learning Models: A Case Study", IEEE Access, Vol. 7, 2019, pp. 177661-177674.

[17] M. Gramajo, L. Ballejos, M. Ale, "Seizing Requirements Engineering Issues through Supervised Learning Techniques", IEEE Latin America Transactions, Vol. 18, No. 07, 2020, pp. 1164-1184.

[18] M. Alkhodari et al. "Screening Cardiovascular Autonomic Neuropathy in Diabetic Patients With Microvascular Complications Using Machine Learning: A 24-Hour Heart Rate Variability Study", IEEE Access, Vol. 9, 2021, pp. 119171-119187.

[19] G. Wang, K. W. Wong, J. Lu, "AUC-Based Extreme Learning Machines for Supervised and Semi-Supervised Imbalanced Classification", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 51, No. 12, 2021, pp. 7919-7930.

[20] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, P. Lestantyo, "Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data", Proceedings of the International Conference on Computer, Control, Informatics and its Applications, Tangerang, Indonesia, 2019, pp. 14-18.

[21] R. Ghorbani, R. Ghousi, A. Makui, 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, 2020, pp. 141066-141079.

[22] E. R. Fernandes, C. P. L. F. A. De Carvalho, X. Yao, "Ensemble of Classifiers Based on Multiobjective Genetic Sampling for Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 32, No. 6, 2020, pp. 1104-1115.

[23] N. Liu, X. Li, E. Qi, M. Xu, L. Li, B. Gao, "A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data", IEEE Access, Vol. 8, 2020, pp. 171263-171280.

[24] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, 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.

[25] K. Anam, H. Ismail, F. S. Hanggara, C. Avian, S. B. Worsito, "Cross Validation Configuration on k-NN for Finger Movements using EMG signals", Proceedings of the International Conference on Instrumentation, Control, and Automation, Bandung, Indonesia, 25-27 August 2021, pp. 17-21.

[26] M. Panda, A. A. A. Mousa, A. E. Hassanien, "Developing an Efficient Feature Engineering and Machine Learning Model for Detecting IoT-Botnet Cyber Attacks", IEEE Access, Vol. 9, 2021, pp. 91038-91052.

[27] J.-G. Choi, I. Ko, J. Kim, Y. Jeon, S. Han, "Machine Learning Framework for Multi-Level Classification of Company Revenue", IEEE Access, Vol. 9, 2021, pp. 96739-96750.

[28] T. Hasanin, T. M. Khoshgoftaar, 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, Los Angeles, CA, USA, 2019, pp. 83-88.

[29] U. R. Salunkhe, S. N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach", Procedia Computer Science, Vol. 85, 2016, pp. 725-732.