DOI: http://dx.doi.org/10.26483/ijarcs.v12i2.6696

ISSN No. 0976-5697

Volume 12, No. 2, March-April 2021

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

WELL-CALIBRATED PROBABILISTIC MACHINE LEARNING CLASSIFIERS FOR MULTIVARIATE HEALTHCARE DATA

Akram Pasha
School of Computer Science and Engineering
REVA University
Bengaluru, India

Latha P. H.
Department of Information Science and Engineering
Sambhram Institute of Technology
Bengaluru, India

Abstract: The healthcare applications frequently collect and store the patient data (mostly multivariate) to examine the history of the treatment and thereby enhance the effectiveness of treatment. The efficient treatment to the patient depends on the performance of the machine learning models used for analytics tasks of patient data. It is convenient to have a machine learning classification model in a healthcare application to predict the probability of an observation belonging to each possible class rather than predicting a class value directly for any disease classification problem. Such predicted probabilities are required to be calibrated to assist the overall support and confidence of any machine learning classification model used in many healthcare applications. In this paper, the predicted probabilities are studied to diagnose and improve the calibration of models used for probabilistic classification. The general performance of selected classification models on the two latest wart skin disease treatment data is also reported.

Keywords: Data Mining, Machine Learning, Classification, Data Analytics, Calibration of Classifiers, Healthcare Systems.

I. INTRODUCTION

In the current era of big data, the technological advancements are boosting the effectiveness in the healthcare applications. Today, doctors are well equipped with the results of advanced analytics performed on the history of patient records to serve the patients effectively. The electronic information about the patients provided to doctors must be increased to enhance the overall effectiveness of the treatment given to the patients. However, having access to the important patterns in the patients' data could be a routine job for any disease diagnostic expert. The diagnostic experts would certainly find it handy to understand the patient's risks in disease through various patterns found in the readings, laboratory test results, race, gender, case history, and socioeconomic standing. Presently, the domain of data analytics has contributed in various spectrums to understand and analyze healthcare data [1-3]. Data analytics has proven to be an effective approach in enhancing the medicinal treatment for the patients [4], facilitating the great advantage to clinicians, to enhance the quality of their expert choices during patient diagnoses. Subsequently, it has contributed to speedy recovery of patients with cost-effective treatment [1-4]. Machine learning has always been the driving force for data analytics, and has been very powerful in analyzing massive data sets that are beyond the normal human capability for analysis [7-9]. Machine learning has the capability of converting the analytical results into the information, suitable for physicians to gain clinical insights that aid them in designing and providing enhanced health care for patients.

The important applications of the proposed study is threefold; it aids the statisticians to explore the behavior of probabilistic classification models towards multivariate data; it equips the physicians with a tool that assists him/her in accurate patient diagnosis based on the probabilistic statistics; and, it aids the ailing patients gain economic medical treatment and rapid recovery.

The following are the major contributions of the work proposed:

- Performs Exploratory Data Analysis (*EDA*) on the multivariate data
- Builds multiple probabilistic classifiers.
- Performs the comparative study of performance of well-calibrated classifiers based on several evaluation metrics.

Let, 'T' be an unseen outcome of the patient undergoing the two wart treatments, 'E' be the set of records showing the results of the patients who have undergone the two wart treatments stored in the form of Comma-Separated-Values, and 'P' be the accuracy of classifying 'T 'based on 'E'. Therefore, the classification problem in the current study can be defined as developing a machine learning model 'M' that gets trained by all the features (called Experience) present in 'E' to predict (a Task) 'T' by improving the accuracy performance 'P' during classification.

The classification (predicting the probabilities) task is an important machine learning task that enables the predictions based on the available data sets referred to as history of treatment. The two data sets of wart treatment therapies chosen in this study is taken from the UCI Machine Learning Repository, contributed by the work of [5].

The rest of the paper is structured as follows. Section-II presents the related work in the field of healthcare data analytics. Section-III presents the detailed framework used in this study. Section-IV presents the experimental setup and the results of exploratory data analysis. Section-V presents detailed visualization and discussion of results. Section-VI concludes the work proposed and further extensions of this work.

II. RELATED WORK

There are many research studies conducted in the area of healthcare data analytics using machine learning. In the work of [7], many basic machine learning algorithms; such as Logistic Regression, KNN, Naïve Bayes and Decision trees

were modelled to predict the heart disease in the patients based on the data recorded from the patients [17][18]. The studies have also been conducted in the past to enumerate the various data mining algorithms for predicting the various diseases. One of the studies reported in [8], review various data mining models and their evaluation methods. The study attempts to determine the most efficient data mining methods used for medical diagnosing purposes. Many studies have been conducted under predictive modelling under the domain of Internet of Things (*IoT*) –based healthcare systems as seen in the work of [9]. Recently, research in the direction of determining the best classifier for five diseases based on the open data sets that are available online was conducted in the work of [10]. In their work, parametric and non-parametric machine learning algorithms were selected and their performance was evaluated. There have been the works in the literature that present the various ways of coupling the advanced technologies with artificial intelligence for effective diagnosis of various diseases [11]. Such systems have contributed in a number of ways to the medicinal communities from enhancing the quality of treatments given to the patients to enhancing the rapid methods of clinical decision making. There are studies conducted for remotely monitoring chronic diseases by collecting and analyzing the physiological data of the patients through the sensors connected to the patients [12]. The advent of big data has also opened many research avenues to the existing data analytics researches to explore the availability of platforms, technologies and open challenges pertinent to huge availability of healthcare data. Some studies conducted in this direction can be seen in the works of [13] [14] [15] [16]. There have been studies reported in the literature incorporating machine learning algorithms to predict the diseases through datasets of treatment and also to deal with the features of datasets, to eventually enhance the performance of the classifiers [20].

Recently, many machine learning classification and regression models were built in the work of [18] [19] to predict the response of treatment for patients having many different types of warts; such as common and/or plantar warts to the cryotherapy and/or Immunotherapy. The literature review shows that exploring various machine learning models on various parameter settings, predicted probabilities, and evaluation of machine learning models based on those parameter settings is still open for research. Additionally, the studies also influence various tools/ technologies used in the overall model development.

III. METHODOLOGY

The Fig. 1 shows the abstract framework of the proposed work. The two datasets of Wart skin disease treatment therapies; Cryotherapy and Immunotherapy are used in this study. The dataset description is as depicted in the tables I and II.

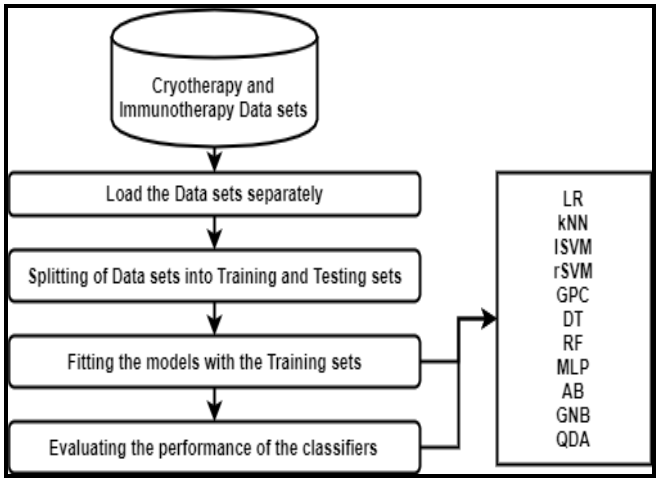


Figure 1. Abstract Framework of the Proposed Work

Table I: Variables in Cryotherapy Treatment Dataset

Variable	*Description of Variable Value Types with their Count*
Sex	Numeric Data 47 records with 'male' and 43 records with 'female'
age	Numeric Data in years. The range of data is 15 to 67
Time	Numeric Data showing 'Time elapsed before treatment' in months. The range of data is 0 to 12
Number_of_Warts	Numeric Data showing the number of warts a patient has. The range of values are 1 to 12
Type	Numeric Data showing the type of warts. The values are '54' for Common warts, '09' for Plantar warts, '27' for both types of warts
Area	Numeric Data showing surface area of wart in millimeters. The range of values are 4 to 750
Result_of_Treatment	Binary Data showing Response_to_treatment. {0,1}

Table II: Variables in Immunotherapy Treatment Dataset

Variable	*Description of Variable Value Types with their Count*
sex	Numeric Data 47 records with 'male' and 43 records with 'female'
age	Numeric Data in years. The range of data is 15 to 67
Time	Numeric Data showing 'Time elapsed before treatment' in months. The range of data is 0 to 12
Number_of_Warts	Numeric Data showing the number of warts a patient has. The range of values are 1 to 12
Type	Numeric Data showing the type of warts. The values are '54' for Common warts, '09' for Plantar warts, '27' for both types of warts
Area	Numeric Data showing surface area of wart in millimeters. The range of values are 4 to 750
Induration	Numeric Data showing diameter of initial test in millimeters. The range of values are 5 to 70
Result_of_Treatment	Binary Data showing Response_to_treatment. {0,1}

The Cryotherapy dataset is the data collected after the Cryotherapy treatment on 90 patients about 6 features. There are 90 rows and 6 columns in the Cryotherapy dataset. The Immunotherapy dataset is the data collected after the Immunotherapy treatment on 90 patients about 7 features. There are 90 rows and 7 columns in the Immunotherapy dataset. In both the dataset the only dependent variable is 'Result of Treatment' which is either positive or negative. The data was collected about the success or failure of two therapies for wart disease treatment on 90 patients. A machine learning classification model for such multivariate data could predict the result of treatment for new patients based on the data that

was used to train the model. Many machine learning models have been researched either on a variety of parameter settings, or on different feature engineering statistics or on a variety of platforms for model development.

In this study, 11 machine learning algorithms (Logistic Regression (*LR*), linear Support Vector Machine (*lSVM*), radial basis function Support Vector Machine (*rSVM*), Gaussian Naïve Bayes (*GNB*), Gaussian Process Classifier (*GPC*), k-Nearest Neighbor (*kNN*), Decision Tree (*DT*), Random Forest (*RF*), Multilayer Perceptron (MLP), Ada Boost (*AB*) and Quadratic Discriminant Analysis (*QDA*)) are employed to investigate the calibration of predicting probabilities of the classifiers. The primary motive of this study is to investigate the fitness and calibration efficiency of classifiers on the two multivariate healthcare datasets.

IV. IMPLEMENTATION

This study uses the open data sets available on UCI Machine Learning Repository [5]. The machine learning algorithms employed in this study are expected to output the predicted probabilities that are interpreted as their confidence levels directly. All the 11 machine learning algorithms are employed using Python-3.6, and scikit-learn v0.20.3 [6], machine learning libraries on Windows platform with a 64-bit computer.

A. *EDA*

The Fig. 2 and Fig. 3 shows the two heat maps that are showing the density of correlations between each of the variables in Cryotherapy and Immunotherapy datasets respectively. It is shown that in both the datasets except 'area' and 'age', all the attributes are found to be highly correlated.

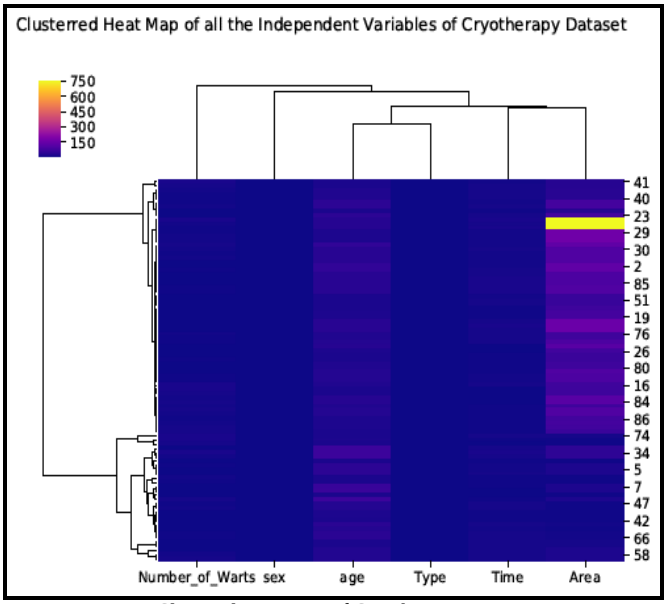


Figure 2. Clustered Heat Map of Cryotherapy Dataset

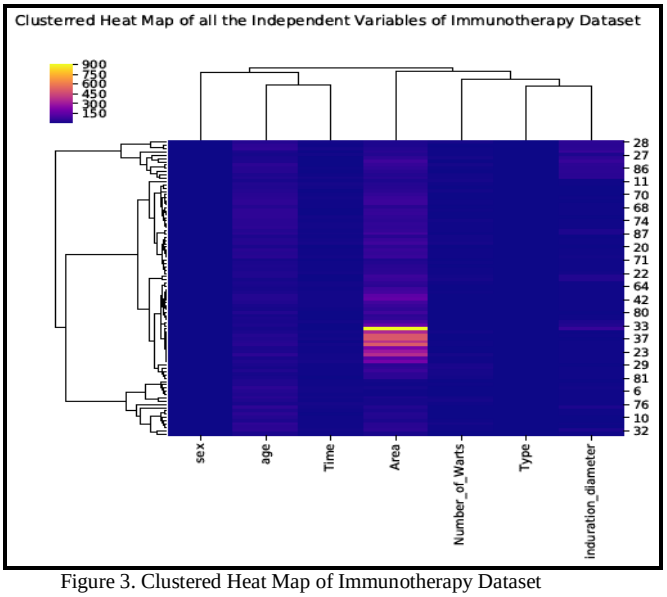


Figure 3. Clustered Heat Map of Immunotherapy Dataset

The Fig. 4 and Fig. 5 are showing the pair plots of Cryotherapy and Immunotherapy datasets respectively, and are used to investigate how each of the attributes are distributed, and to investigate whether the datasets are linearly separable. The pair plots show that only the attributes: 'age', time', and 'number of warts' are having slight uniformity in distribution.

Within the context of the problem being solved, there are two major objectives, first is to get the most efficient classification model for the two datasets, the second is to investigate the calibration of the probabilistic classifiers to improve the confidence level in the accuracy of classification. The attributes that are showing low in the heat maps are contributing to the classification accuracy of the classification model.

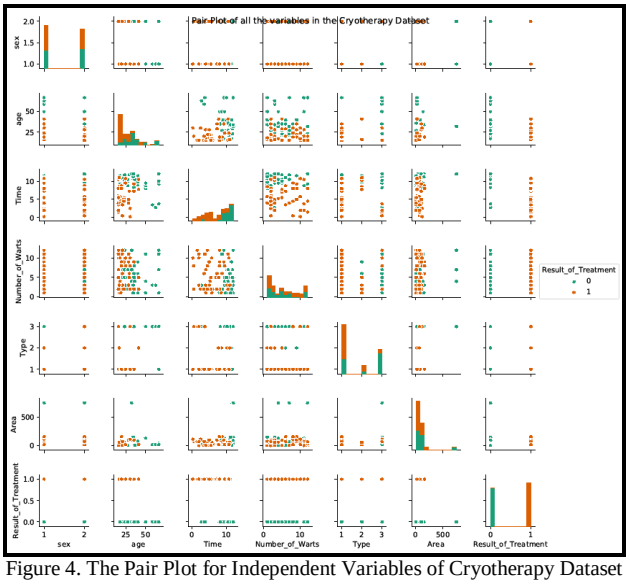


Figure 4. The Pair Plot for Independent Variables of Cryotherapy Dataset

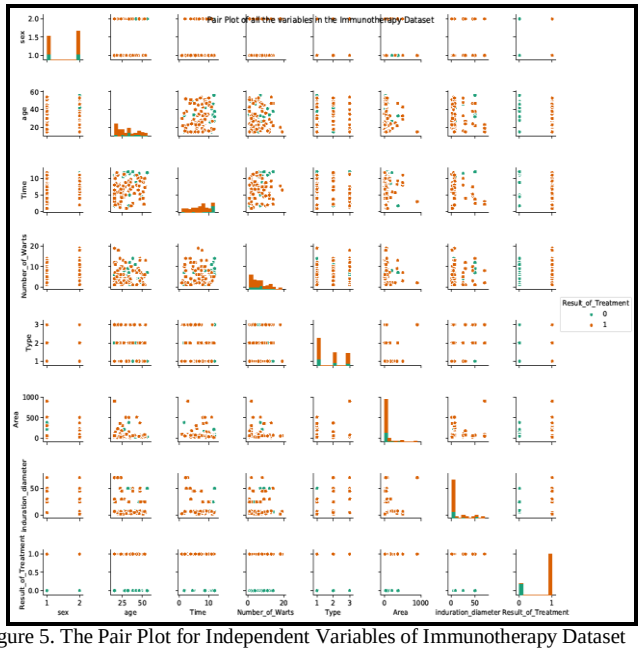
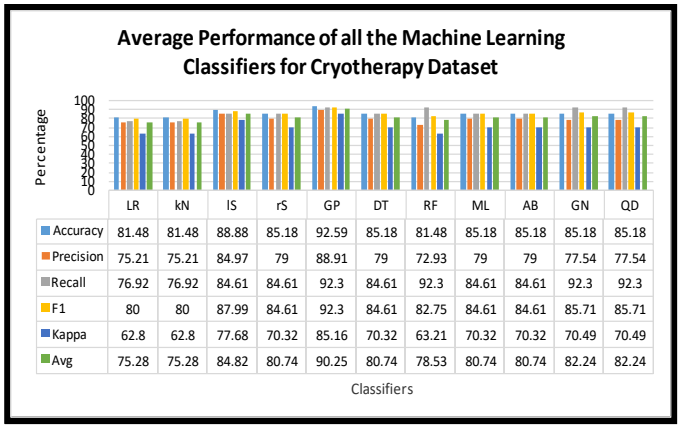


Figure 5. The Pair Plot for Independent Variables of Immunotherapy Dataset

V. EXPERIMENTAL RESULTS AND DISCUSSION

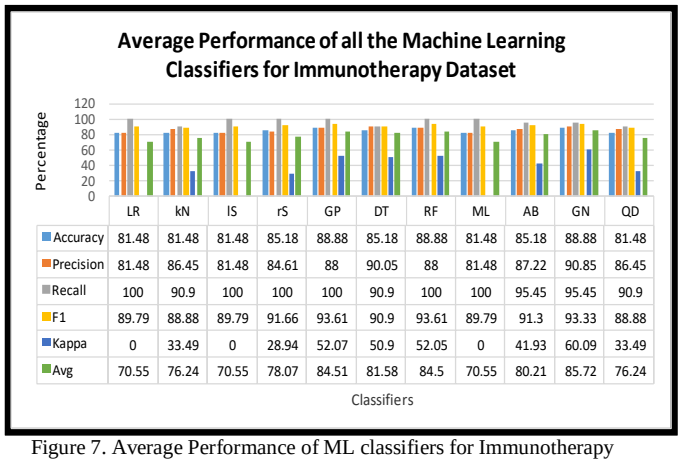
The Fig. 6 shows the average performance of all the 11 classifiers on the Cryotherapy dataset. The GNB classifier outperformed all the other classifiers giving 85.72%. This shows that the nature of data used to train any binary classifier is bound by the probabilistic characteristic of a GNB classifier. The GP classifier and the RF classifier are found to be almost equivalent to GNB classifier in their performance giving 84.51% and 84.50%.



	LR	kN	lS	rS	GP	DT	RF	ML	AB	GN	QD
Accuracy	81.48	81.48	88.88	85.18	92.59	85.18	81.48	85.18	85.18	85.18	85.18
Precision	75.21	75.21	84.97	79	88.91	79	72.93	79	79	77.54	77.54
Recall	76.92	76.92	84.61	84.61	92.3	84.61	92.3	84.61	84.61	92.3	92.3
F1	80	80	87.99	84.61	92.3	84.61	82.75	84.61	84.61	85.71	85.71
Kappa	62.8	62.8	77.68	70.32	85.16	70.32	63.21	70.32	70.32	70.49	70.49
Avg	75.28	75.28	84.82	80.74	90.25	80.74	78.53	80.74	80.74	82.24	82.24

Figure 6. Average Performance of ML classifiers for Cryotherapy Dataset

The Fig. 7 shows the average performance of all the 11 classifiers on the Immunotherapy dataset. The GP classifier outperformed all the other classifiers with 90.25%. The results show that the nature of data used to train binary classifiers is bound by the Gaussian distribution of data that is bound by the algorithm. The lSVM was slightly better giving 84.82% compared to GNB and QDA classifiers' performance giving 82.24%.



	LR	kN	lS	rS	GP	DT	RF	ML	AB	GN	QD
Accuracy	81.48	81.48	81.48	85.18	88.88	85.18	88.88	81.48	85.18	88.88	81.48
Precision	81.48	86.45	81.48	84.61	88	90.05	88	81.48	87.22	90.85	86.45
Recall	100	90.9	100	100	100	90.9	100	100	95.45	95.45	90.9
F1	89.79	88.88	89.79	91.66	93.61	90.9	93.61	89.79	91.3	93.33	88.88
Kappa	0	33.49	0	28.94	52.07	50.9	52.05	0	41.93	60.09	33.49
Avg	70.55	76.24	70.55	78.07	84.51	81.58	84.5	70.55	80.21	85.72	76.24

Figure 7. Average Performance of ML classifiers for Immunotherapy Dataset

Table III. Average Performance of all the Machine Learning Classifiers

	Accuracy		Precision		Recall		F1		Kappa		Avg
	C	I	C	I	C	I	C	I	C	I	
LR	**81.48**	**81.48**	**75.21**	**81.48**	**76.92**	**100.0**	**80.0**	**89.79**	**62.80**	**0.0**	**72.91**
kN	**81.48**	**81.48**	**75.21**	**86.45**	**76.92**	**90.90**	**80.0**	**88.88**	**62.80**	**33.49**	**75.76**
lS	**88.88**	**81.48**	**84.97**	**81.48**	**84.61**	**100.0**	**87.99**	**89.79**	**77.68**	**0.0**	**77.68**
rS	**85.18**	**85.18**	**79.00**	**84.61**	**84.61**	**100.0**	**84.61**	**91.66**	**70.32**	**28.94**	**79.41**
GP	**92.59**	**88.88**	**88.91**	**88.0**	**92.30**	**100.0**	**92.30**	**93.61**	**85.16**	**52.07**	**87.38**
DT	**85.18**	**85.18**	**79.00**	**90.05**	**84.61**	**90.90**	**84.61**	**90.90**	**70.32**	**50.90**	**81.16**
RF	**81.48**	**88.88**	**72.93**	**88.0**	**92.30**	**100.0**	**82.75**	**93.61**	**63.21**	**52.05**	**81.52**
MP	**85.18**	**81.48**	**79.0**	**81.48**	**84.61**	**100.0**	**84.61**	**89.79**	**70.32**	**0.0**	**75.64**
AB	**85.18**	**85.18**	**79.0**	**87.22**	**84.61**	**95.45**	**84.61**	**91.30**	**70.32**	**41.93**	**80.48**
GN	**85.18**	**88.88**	**77.54**	**90.85**	**92.30**	**95.45**	**85.71**	**93.33**	**70.49**	**60.09**	**83.98**
QD	**85.18**	**81.48**	**77.54**	**86.45**	**92.30**	**90.90**	**85.71**	**88.88**	**70.49**	**33.49**	**79.24**

The Table III shows the performance of all the 11 classifiers on both the datasets collectively. The GP algorithm outperformed all the other classifiers giving 87.38%. The GNB classifier was slightly better with 83.98% compared to Random Forest classifier with 83.98%.

The figures (Fig. 8 to Fig. 11) are the calibration plots of all the 11 classifiers that are employed on the two datasets.

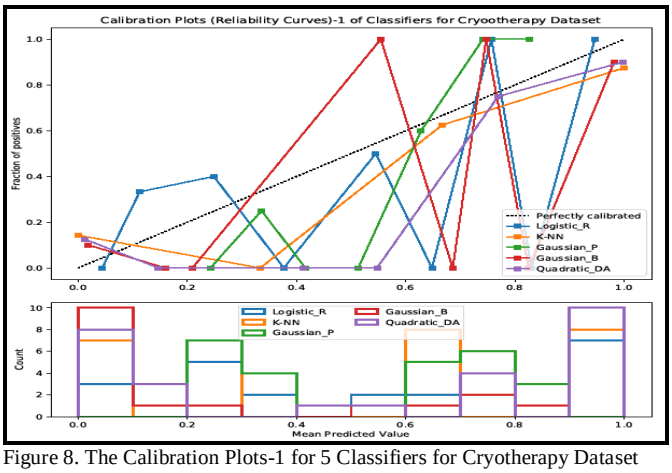


Figure 8. The Calibration Plots-1 for 5 Classifiers for Cryotherapy Dataset

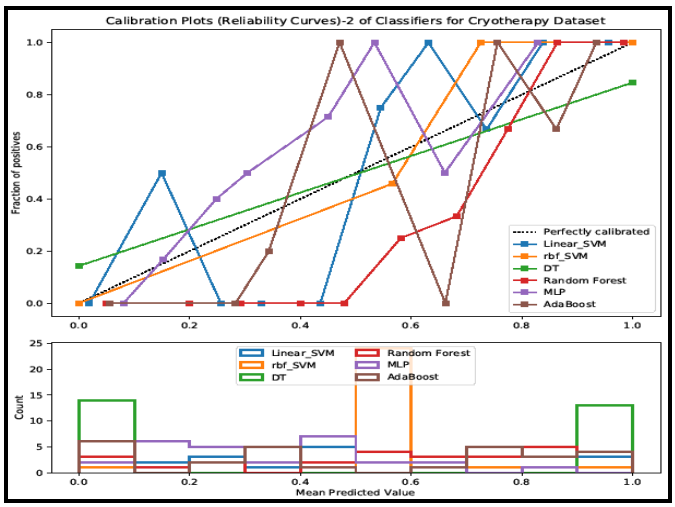


Figure 9. The Calibration Plots-2 for next 6 Classifiers for Cryotherapy Dataset

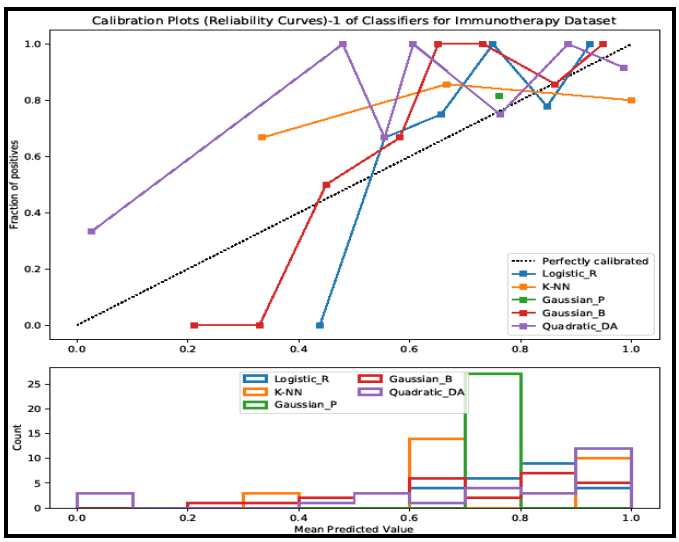


Figure 10. The Calibration Plots-1 for 5 Classifiers for Immunotherapy Dataset

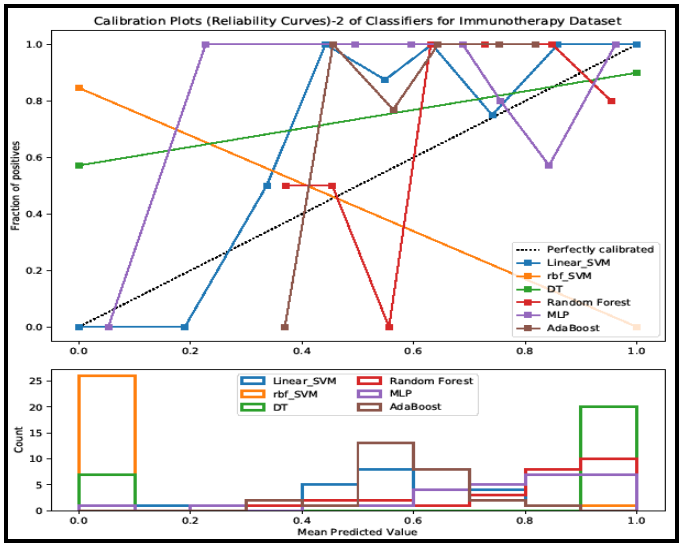
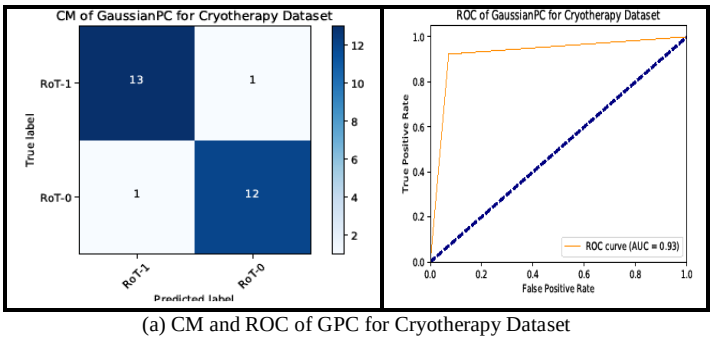


Figure 11. The Calibration Plots-2 for next 6 Classifiers for Immunotherapy Dataset

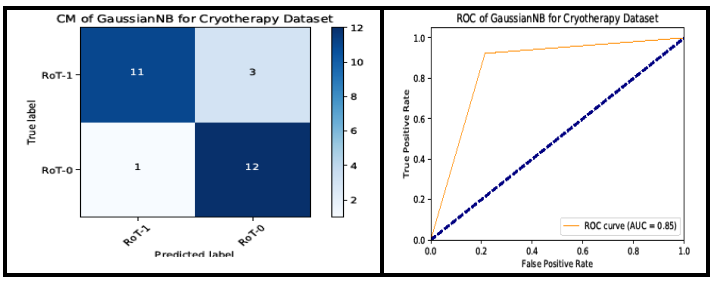
The Fig. 12 shows the confusion matrices and the ROC curves of the well calibrated classifiers for the two data sets. It can be seen in the figure 8 and 10 that GP and GNB probabilistic classification models are well calibrated classifiers for which the output of the predicted probabilities are directly interpreted as a confidence level for both Cryotherapy and Immunotherapy datasets. For instance, a well calibrated GP classifier is classifying the observations, such that among the observations to which it gave a predicted probability value close to 86% as shown in figure 6, is belonging to the positive class as seen in the figure 12(a) with the corresponding confidence levels as shown in the figures 11 and 12(b) through confusion matrix and ROC curves for Cryotherapy dataset. Similar observations can be seen in the figures 7, 12(c) and 12(d) for Immunotherapy dataset.

VI. CONCLUSION AND FUTURE ENHANCEMENT

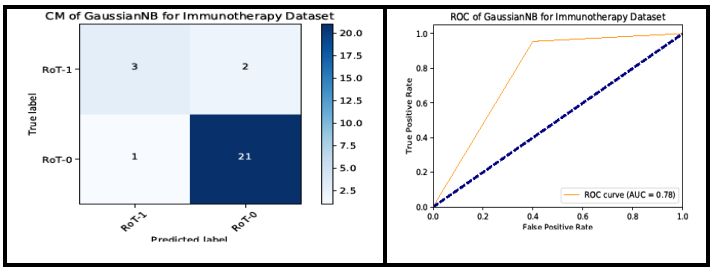
The healthcare data analytics have provided unprecedented advantages in tendering effective medical diagnosis and treatment of diseases. In the current research study, the combination of linear and non-linear algorithms; such as Logistic Regression, SVMs, Decision Tree, Random Forest, Gaussian Naïve Bayes, Gaussian Process, Adaboost and Quadratic Discriminant Analysis classification models were employed. The property of probabilistic classification models to become a well-calibrated classifier is investigated experimentally on the two healthcare datasets of wart treatment methods. The work proposed does not consider parameter tuning of the classification models employed. The future enhancement could be to tune the parameters and /or hyper parameters of the classifiers to improve the further performance, and also choose the multidimensional big datasets from different problem domains. The investigation of such datasets, on a distributed computing platform with parameter tuning can also be one of the extensions of the work.



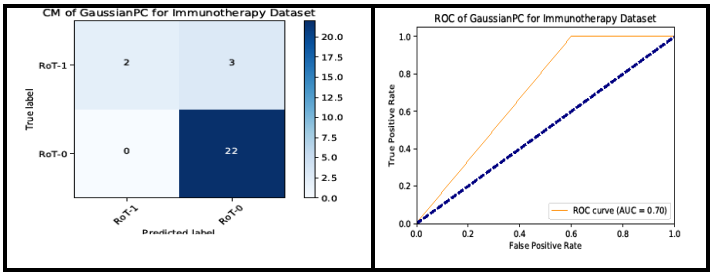
(a) CM and ROC of GPC for Cryotherapy Dataset



(b) CM and ROC of GNB for Cryotherapy Dataset



(c) CM of ROC of GNB for Immunotherapy Dataset



(d) CM and ROC of GPC for Immunotherapy Dataset

Figure 12. Classification Performance of the selected classifiers

VII. ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the management and all the personnel of both REVA University, and Sambhram Institute of Technology, Bengaluru, India for providing all the support for carrying out this work.

VIII. REFERENCES

[1] Ismail A, Shehab A, El-Henawy IM. Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations. InSecurity in Smart Cities: Models, Applications, and Challenges 2019 (pp. 27-45). Springer, Cham.

[2] Ottenbacher KJ, Graham JE, Fisher SR. Data Science in Physical Medicine and Rehabilitation: Opportunities and Challenges. Physical Medicine and Rehabilitation Clinics. 2019 Mar 2.

[3] Milenkovic MJ, Vukmirovic A, Milenkovic D. Big data analytics in the health sector: challenges and potentials. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 2019 Mar 19.

[4] Delen D, Davazdahemami B, Eryarsoy E, Tomak L, Valluru A. Using predictive analytics to identify drug-resistant epilepsy patients. Health informatics journal. 2019 Mar 12:1460458219833120.

[5] Khozeimeh F, Alizadehsani R, Roshanzamir M, Khosravi A, Layegh P, Nahavandi S. An expert system for selecting wart treatment method. Computers in biology and medicine. 2017 Feb 1; 81:167-75.

[6] Pedregosa et al.,Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

[7] Arjaria SK, Rathore AS. Heart Disease Diagnosis: A Machine Learning Approach. InAdvanced Classification Techniques for Healthcare Analysis 2019 (pp. 161-181). IGI Global.

[8] Ghorbani R, Ghousi R. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. International Journal of Data and Network Science. 2019;3(2):47-70.

[9] Sharma M, Singh G, Singh R. An Advanced Conceptual Diagnostic Healthcare Framework for Diabetes and Cardiovascular Disorders. arXiv preprint arXiv:1901.10530. 2019 Jan 13.

[10] Singh AK. A Comparative Study on Disease Classification using Machine Learning Algorithms. Available at SSRN 3350251. 2019 Mar 11.

[11] Vashistha R, Yadav D, Chhabra D, Shukla P. Artificial Intelligence Integration for Neurodegenerative Disorders. InLeveraging Biomedical and Healthcare Data 2019 Jan 1 (pp. 77-89). Academic Press.

[12] Syed L, Jabeen S, Manimala S, Elsayed HA. Data Science Algorithms and Techniques for Smart Healthcare Using IoT and Big Data Analytics. InSmart Techniques for a Smarter Planet 2019 (pp. 211-241). Springer, Cham.

[13] Kari V, Amalanathan GM. Synthesis of Classification Models and Review in the Field of Machine Learning. InAdvanced Classification Techniques for Healthcare Analysis 2019 (pp. 18-51). IGI Global.

[14] Razzak MI, Imran M, Xu G. Big data analytics for preventive medicine. Neural Computing and Applications.1-35.

[15] Bucholc M, Ding X, Wang H, Glass D, Wang H, Prasad G, Maguire L, Bjourson A, McClean P, Todd S, Finn D. A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. bioRxiv. 2019 Jan 1:573899.

[16] Dehkordi SK, Sajedi H. Prediction of disease based on prescription using data mining methods. Health and Technology. 2019 Jan 24; 9(1):37-44.

[17] Gambhir S, Kumar Y, Malik S, Yadav G, Malik A. Early Diagnostics Model for Dengue Disease Using Decision Tree-Based Approaches. InPre-Screening Systems for Early Disease Prediction, Detection, and Prevention 2019 (pp. 69-87). IGI Global.

[18] Ramana BV, Boddu RS. Performance Comparison of Classification Algorithms on Medical Datasets. In2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) 2019 Jan 7 (pp. 0140-0145). IEEE.

[19] Ghiasi MM, Zendehboudi S. Decision tree-based methodology to select a proper approach for wart treatment. Computers in Biology and Medicine. 2019 Apr 4.

[20] Pasha, Akram, and P. H. Latha. "Bio-inspired dimensionality reduction for Parkinson's disease (PD) classification." Health information science and systems 8.1 (2020): 1-22.