Machine Learning Based Early Detection System of Cardiac Arrest

Ji-Han Liu
Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan
d03944005@ntu.edu.tw

Hsiao-Ko Chang
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
d06922027@ntu.edu.tw

Cheng-Tse Wu
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
d02922011@ntu.edu.tw

Wee Shin Lim
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
leolim3092@csie.ntu.edu.tw

Hui-Chih Wang
Department of Emergency Medicine
National Taiwan University Hospital
Taipei, Taiwan
ticoer@ntuh.gov.tw

Jyh-Shing Roger Jang
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
jang@csie.ntu.edu.tw

Abstract

Target—Most of the Cardiac Arrest (CA) cases are preventable because the CA patients usually had abnormal clinical signs or symptoms before their suffering from CA. In general, the appropriate steps of Cardiopulmonary resuscitation (CPR) for CA patients will increase the patients' survival rate and reduce the consequent medical expenses. Accordingly, we propose a system and the related methods for detecting CA before the CPR event occurred earlier and it is not only assisting physicians to early diagnose of CA and immediately warning but also improving the medical quality.

Methods—In this study, the raw dataset is collected from the electronic health records (EHRs) of the adult patients (age ≧20 years) who visited emergency department (ED) and stayed in the emergency detention area for more than 6 hours during January 2014 to December 2015, and it is provided by National Taiwan University Hospital (NTUH). We perform the data preprocessing and cleaning for the dataset using a resampling technique to balance the data amount of CPR and Non-CPR patients of the dataset, and then we construct a sliding window and apply several classifiers for model training and reducing the possible overfitting problem. Additionally, we use the measures such as the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) to comparative evaluate the performance of our models built.

Results—Our approach avoids the problems of dataset imbalance and possible overfitting effectively. The performance among classifiers selected show that the best one is random forest (RF) when CPR event happened, but the better ones are Logistic Regression and LSTM than the remaining classifiers and close to that of RF during 1 to 4 hours before the CPR time.

Conclusion—We have the contribution in predicting CA and CPR event before it occurred around 3 to 3.5 hours in advance, and it is similar to that of the state-of-the-art such Early Warning Score (EWS) with around 3.5 hours. In addition, avoiding the problem of dataset imbalanced may effectively improve the accuracy of predicting CA as well. Accordingly, it helps to assist the emergency clinical physicians to achieve the hospital's quality management including the clinical or medical resources allocation.

Keywords—Cardiac Arrest, CPR, resampling, sliding window, overfitting, machine learning, deep learning, early detection, prediction

I. INTRODUCTION

Cardiac arrest (CA) is a sudden loss of blood flow resulting from the failure of the heart to effectively pump, and the evidence shows that there are 80% of the patients with CA will show signs of deterioration in the 8 hours before cardiac arrest [1]. In addition, cardiopulmonary resuscitation (CPR) is an emergency procedure that developed for the person who is in sudden suffered CA or respiratory arrest [2, 3]. Therefore, it is very important to detect CA earlier by developing or improving an early detecting or warning system to avoid applying CPR to the patients who stayed in the **e**mergency detention area for more than 6 hours long due to the insufficient medical resources or their critical situation shall be observed for a fixed duration.

In addition, a great amount of electronic health records (EHRs) have become available in recent years, and the features regarding health dataset include age, gender, weight, triage, vital signs, etc. In the case of having the imbalanced data in the raw dataset, it causes the inefficient steps and duplication in repeating or interpolating the missing data and results in the biases on feature selection. Therefore, in our parallel but previously published paper [4], we adopts the mechanisms such as the undersampling to solve the problem of imbalanced dataset, the shifting windows with and without overlap respectively, the AUROC curves and F_3 score to measure the performance.

So far there are a lots of classifiers such as Random Forest [5], Naïve Bayes [6, 7], AdaBoost [8, 9], C4.5 (i.e., Decision Tree) [10], logistic regression (LR), CNN, Long Short-Term Memory (LSTM) [11] for training and test the

IEEE978-1-7281-4666-9/19/$31.00 ©2019 IEEE

models for prediction. It's believed that Long Short-Term Memory (LSTM) is an effective deep learning method for training the model for the data having the temporal property.

The major contributions of our work are:

- We preprocess the imbalanced dataset by resampling techniques and use some vital signs as important features to detect CA earlier and rapidly.
- We design and apply the sliding window without overlap in the model training stage to overcome the possible over-fitting problem that the accuracy is dramatically lower than that generated or outputted training model. Additionally, we solve the shortage problem of training data by determining the reasonable fixed length of sliding window without overlap as well.
- Comparing to our parallel study which is published previously [11], we provide another solution of imbalanced dataset, and we take only 8 hours at most before the CPR or the leave time as observation time interval. For the shorter observation time that satisfying the real condition for predicting the incoming CPR event, we evaluate that the performance of this study is acceptable because it reveals that the similar results of AUROC between this study and the one published previously.

The remainder of this paper is organized as follows: We will briefly review the related works in Section 2, describe the proposed methods in Section 3, depict experiments in Section 4, report results from experiments in Section 5, and finally conclude this paper in Section 6.

II. Related Work

Data in datasets are commonly imbalanced in the medical field when the class distributions are highly imbalanced, and the performance would be terribly affected by the imbalanced data [12]. More precisely to illustrate, the accuracy outputted by the training model is related to the feature selected. Class imbalance problem occurs when a dataset is dominated by a major class or by classes which have significantly more instances than the other rare or minority classes in the data. For the two-class case, without loss of generality, one assumes that the minority or rare class is the positive class, and the majority class is the negative class. Often the minority class is very infrequent, such as 1% of the dataset. If one applies the most traditional classifiers on the dataset, they are likely to predict everything as negative (the majority class). However, typically, people have more interest in learning about rare classes. For example, applications such as medical diagnosis prediction of rare but important disease, such as cancer, where it is common to have fewer cancer patients than healthy patients.

To overcome the problem of imbalance, some approaches can be implemented in the preprocessing stage. Data sampling, in which the training samples are modified in such a way as to produce a more or less balanced class distribution that allows classifiers to improve their performance. Traditional data sampling techniques are undersampling which creates a subset of the original dataset by eliminating instances, oversampling which creates a superset of the original dataset by replicating some instances or creating new instances from existing ones to balance the skewed class ratio, and hybrid method, i.e. resampling, that includes both undersampling and oversampling to the dataset. There are some methods to oversample, resample or under-sample a dataset used in the typical classification problem, and the common techniques are known as Synthetic Minority Over-sampling Technique (SMOTE) [13], Borderline-SMOTE [14], Adaptive Synthetic Sampling Approach (ADASYN) [15], Random Sampling and stratified sampling. SMOTE is one of the most employed resampling techniques, where the minority class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbors.

Moreover, for a binary class problem, the imbalance degree of a class distribution can be denoted by the ratio of the sample size of the small class to that of the prevalent class [16]. In practical applications, the ratio can be as drastic as 1:100, 1:1000, or even larger [17]. Furthermore, one research is conducted to explore the relationship between the class distribution of a training dataset and the classification performances of decision trees. Therefore, the mentioned studies indicate that a relatively balanced distribution usually attains a better result, and a ratio as low as 1: 35 can make some methods inadequate for building a good model [16].

In our parallel but previously published paper [13], it adopts the mechanisms such as the undersampling to solve the problem of imbalanced dataset, the shifting windows with and without overlap respectively, the AUROC curves and F_3 score to measure the performance. The functionality or the necessity of taking F_3 score as the measure for the application is that if our concern is recall .

III. Methods

Fig. 1 shows the workflow of our proposed method. In this study, we collect the dataset from EHRs of the adult patients visited and stayed at NTUH's ED for more than 6 hours from 2014 to 2015, and we de-identify all patients' information before the analysis. Consequently, we identify 124 CA positive patients and 43,445 CA negative patients in the ED, perform a 3-fold cross-validation to raw dataset that we take two parts and one part of them as the training and test dataset respectively, and take the resampling approach to add similar data of underrepresented class to balance the class ratio. Explaining in detail, for the training we adjust the oversampling rate and the undersampling rate to 1:1, for test we adjust the undersampling rate of raw dataset to 1:10, and finally we get 9270 CPR patients and 9270 non-CPR patients, and the ratio is eventually 1:1.

Moreover, based on the Early Warning Score (EWS) [1, 18] or National Early Warning Score (NEWS) [19] commonly used in the medical field, we obtain the observation index of vital signs by the measurement of the NEWS's score sheet, and thus the features used in this study are age, gender, GCS, pulse, body temperature, systolic blood pressure, diastolic blood pressure, blood oxygen concentration (spo2), respiration rate, etc. Accordingly, we select the features having the property of

time series matched with that of NEWS for comparison. Therefore, the important features regarding the vital signs in this study include pulse, body temperature (i.e., BT), systolic, spo2, and respiration rate, and they are collected multiple times at irregular intervals during the patients' registered time to the discharge time or the CPR time.

Again, as shown in Fig. 1, we design the fixed length sliding window without overlap for preventing the decrease of accuracy due to the duplicated use of the data augmented for the missing data in the step of model training shown in Fig. 1 and Fig. 2.

During 'Test' phase, we deploy the dataset for CPR, and performing the steps of feature extraction and predicting and then outputting the prediction report for emergency clinical physicians or examinee.

Fig. 2 illustrates the concepts of the sliding window proposed. During the observing time interval (e.g., 8 hours) that the patient stayed in the emergency detention area of ED, the sliding window slides from a sliding time, 0,1,… to n, n=4 if length of the sliding window is 2 hours, distancing from the leave or discharge time or the CPR time (i.e., 0 hour) that CPR event occurred for patient 1 and 2 respectively.

Again, as shown in Fig. 2, for the case of 1 to 4 hours that of the patient 2, the sliding window with length m=2 hours slides from a sliding time 0 to 4 hours before the CPR time (i.e., 0 hour) or leave time.

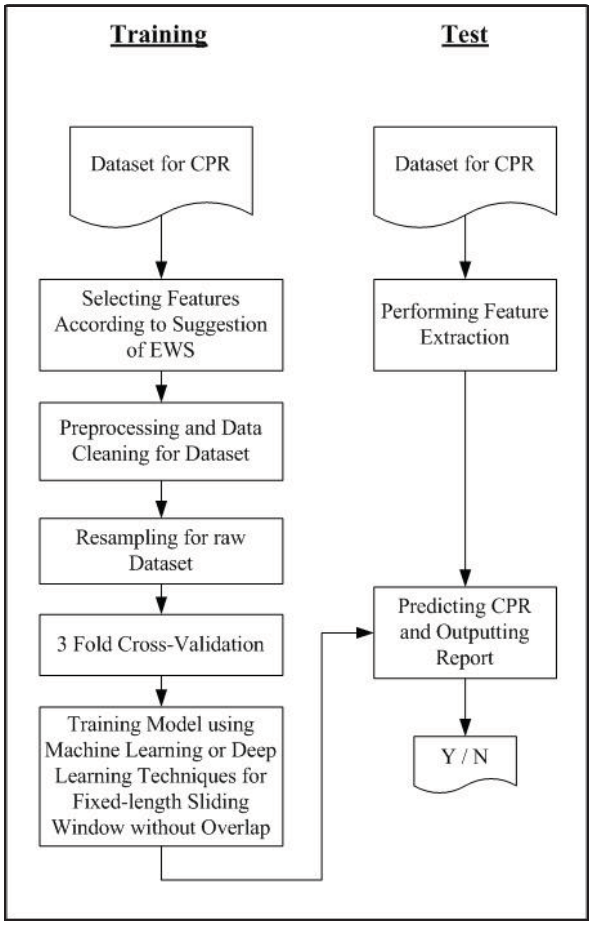

Fig. 1. Flowchart of proposed method

We take 8 hours before the CPR event as the observation period, and every 2 hours of them as an observation interval of a sliding window to meet the rationality of clinical emergency medicine. That is, they usually take 8 hours as nursing shift time and thus there are fixed and continuous measurement of vital signs, 6 hours is the normal administration time, and 4 hours is half of the nursing work period. Additionally, the sliding window is composed of a sequence of partition, and each partition we collects a number of EHRs containing the vital signs for detecting whether a patient will be applied CPR several hour later or not. Therefore, we define the length of the sliding window, m hour(s), i.e., m=2 hours. When the sliding window is slide without overlap, the old partition is disregarded and a new partition containing a set of newly collected EHRs are appended to the window.

Regarding the case of using sliding window without overlap for predicting whether the patient will apply CPR several hours later, the patients' sequences are right aligned and then we construct the fixed length sliding windows for every patient. When an emergency patient visit the ED, medical staffs measure the patient's information such as the vital signs for us to analyze information of every patient. For CA positive patients such as patient 2, the window #1 is a sliding window that nearest the CPR time. In other work, maybe we can adopt the variable length for the window according to the degree it nears the CPR time or other reference factors is work as well.

In this study, we take the same features according to the prior research, NEWS, as a comparison, and we use K-Nearest Neighbor (KNN) and the leave-one-out (LOO) test for F_1 score estimation. For measurement, we apply the normalization function to $F(X_i)$ shown in (1) where X_i representing to a i_{th} value of raw data adjust the values of the features, and the range of these normalized values is between 0 and 1.

$$F(x_i) = \frac{x_i - \min(\forall x_i)}{\max(\forall x_i) - \min(\forall x_i)} \quad (1)$$

Moreover, we use the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC) and F_β-Score shown in (2) as evaluation metrics, where β is 1, 2 and 3 commonly.

$$F_\beta(\text{Score}) = \frac{(\beta^2+1) * precisioin * recall}{\beta^2 * precision + recall} \quad (2)$$

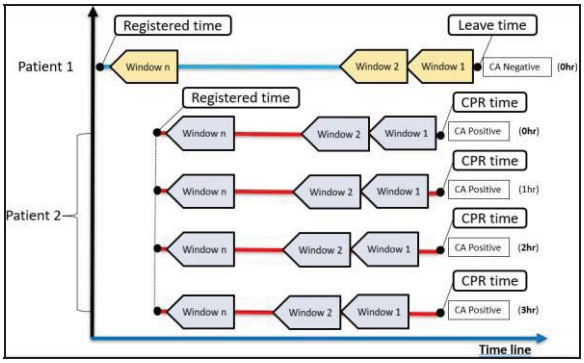

Fig. 2. Diagram of sliding window without overlap (0 hour)

IV. EXPERIMENTS

The machine learning is implemented using the scikit-learn package in Python [20], and the neural networks are implemented in Keras [21] with TensorFlow [22] as the backend engine. Regarding the machine learning and deep learning, we use and compare the performance of AdaBoost, Random Forest algorithm, Naïve Bayes, C4.5 (Decision Tree), CART, Logistic Regression, LSTM, CNN and CNN+LSTM. In this study, we implemented nine different models and classifiers, and used RF as a benchmark, because RF is better in the related work such as EWS. Again, we also use new technologies such as DL (LSTM, CNN, CNN+LSTM) for comparison.

In our experiments, the vital signs of the ordinary patients were measured at least 3 times per day manually by the medical staffs that work at the ED and exchanged in returns, and thus we designed to detect CA using the sliding window which length is 2 hours. As mentioned above, we obtain the observation index of vital signs based on the measurement of the NEWS's score sheet. The features having the property of time series matched with that of NEWS regarding the vital signs in this study include pulse, body temperature (i.e., BT), systolic, spo2, and respiration rate.

Consequently, based on the workflow shown in Fig. 1, we perform the 3-fold cross-validation to raw dataset with 124 CA positive patients and 43,445 CA negative patients in the ED and resampling to balance the class ratio to get 9270 CPR patients and 9270 non-CPR patients with the ratio 1:1. Again, as shown in Fig. 2, the sliding window with length m=2 hours slides from a sliding time 0 to 4 hours before the CPR time (i.e., 0 hour) or leave time for each patient. After the stage of model training, we perform test with the steps of feature extraction and prediction, and obtain AUROC and AUPRC (not shown) for F_1 score estimation. For Clarify, we list the value of F_1 score for all the classifier in Table I.

V. RESULTS

In this section, we present the results for the performance of the case using the sliding window without overlap. Fig. 3(a) to 3(e) illustrate the AUROC curves of the sliding window without overlap that detecting 0 to 4 hours before CPR time or leave time respectively, and Fig. 4 shows the bar chart for the results of AUROC curves.

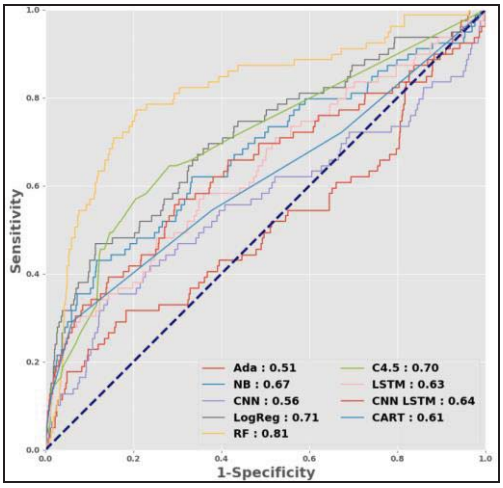

Fig. 3(a). AUROC (0 hour)

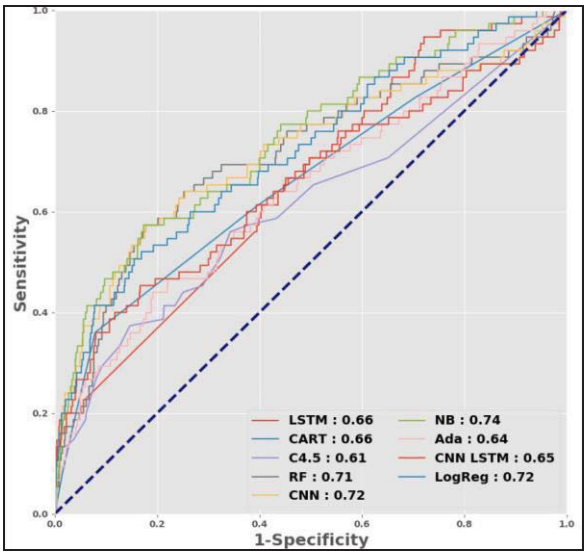

Fig. 3(b). AUROC (1 hour)

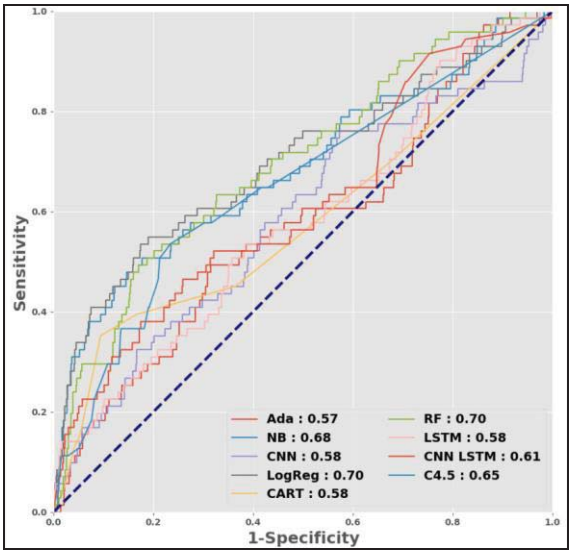

Fig. 3(c). AUROC (2 hours)

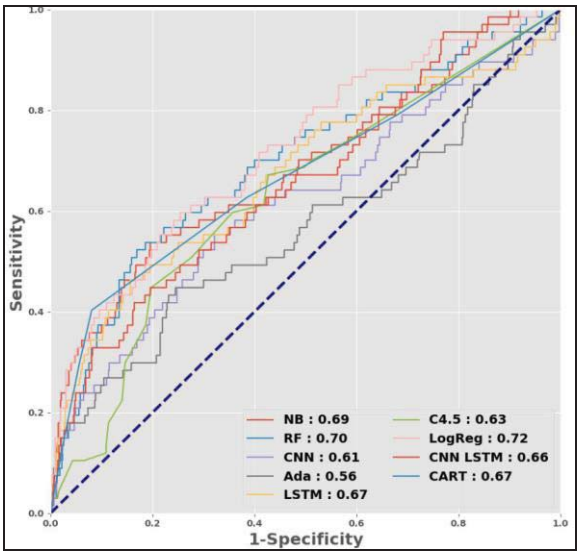

Fig. 3(d). AUROC (3 hours)

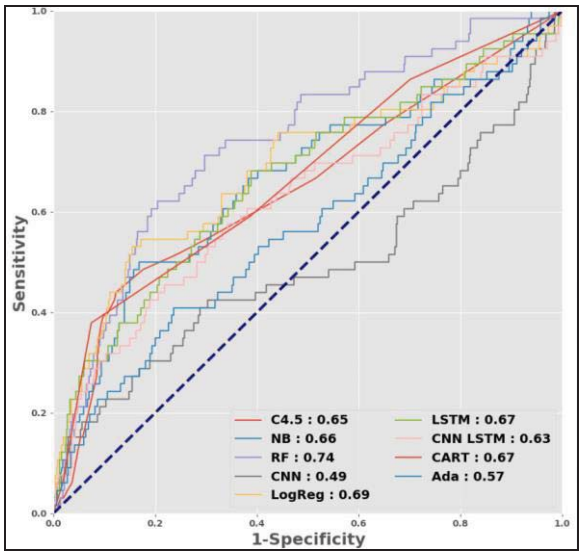

Fig. 3(e). AUROC (4 hours)

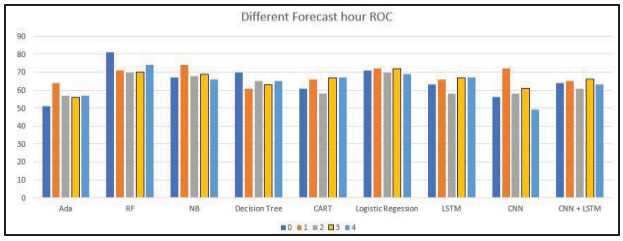

Fig. 4. Comparison of AUROC value (%) of different forecast hour

Again, as shown in Fig. 3(a) to 3(e) and 4, the best performance among different classifiers in this experiment is random forest (RF) when CPR event happened (0 hour, AUROC value: 0.81) and the time 4 hours before that (AUROC value: 0.74) represented by AUROC, but Logistic Regression (AUROC value: 0.69), LSTM (AUROC value: 0.67) and CNN+LSTM (AUROC value: 0.63) are better than other classifier and close to that of RF when 4 hours before the CPR event happened.

Table I shows the results of F_1 score, and Fig. 5(a) to 5(e) shows the F_1 score for 0 to 4 hours for the sliding window without overlap. The best accuracy of prediction is RF (0,4 hour, F_1 score: 031, 0.23) consistenting with prior research and general understanding. In addition, the better one is Logistic regression (1 to 4 hours, F_1 score: 0.32, 0.28, 0.3, 0.27 respectively) in this experiment, and that of LSTM (4 hours, F_1 score: 0.22) is close to them as well.

Table I. RESULTS OF F_1 SCORE (0 TO 4 HOURS)

Classifier	F1_0h	F1_1h	F1_2h	F1_3h	F1_4h
Ada	0.13	0.21	0.13	0.19	0.15
RF	**0.31**	0.27	0.21	0.22	0.23
NB	0.25	0.3	0.26	0.27	0.2
Decision Tree(C4.5)	0.23	0.2	0.17	0.16	0.18
CART	0.2	0.19	0.15	0.21	0.19
Logistic Regression	0.29	**0.32**	**0.28**	**0.3**	**0.27**
LSTM	0.21	0.3	0.18	0.27	0.22
CNN	0.21	0.24	0.15	0.18	0.16
CNN+LSTM	0.24	0.3	0.16	0.22	0.18

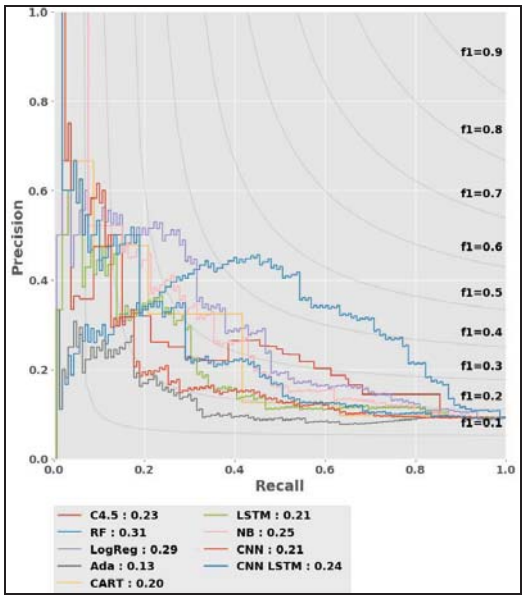

Fig. 5(a). F_1 score (0 hour)

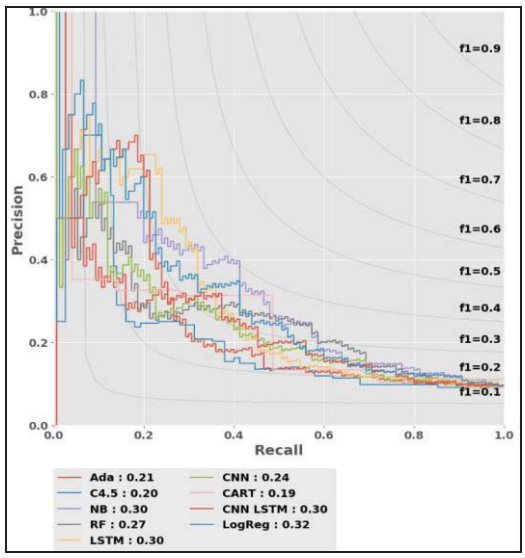

Fig. 5(b). F_1 score (1 hour)

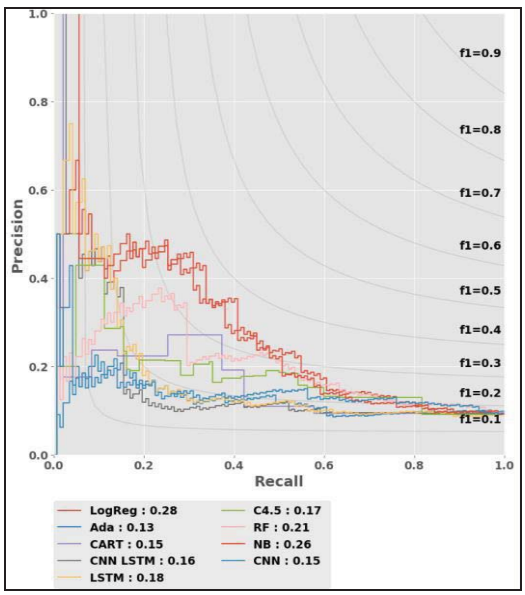

Fig. 5(c). F_1 score (2 hours)

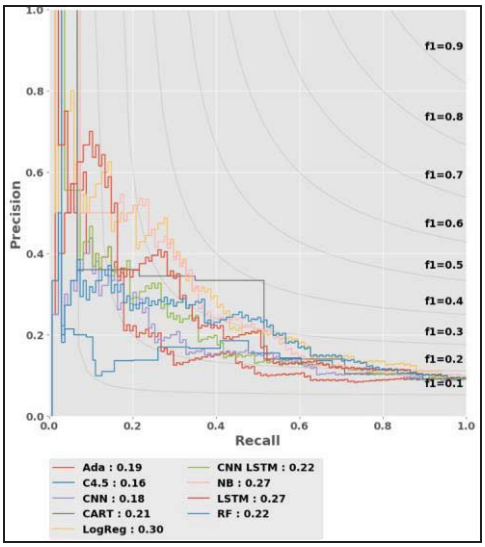

Fig. 5(d). F_1 score (3 hours)

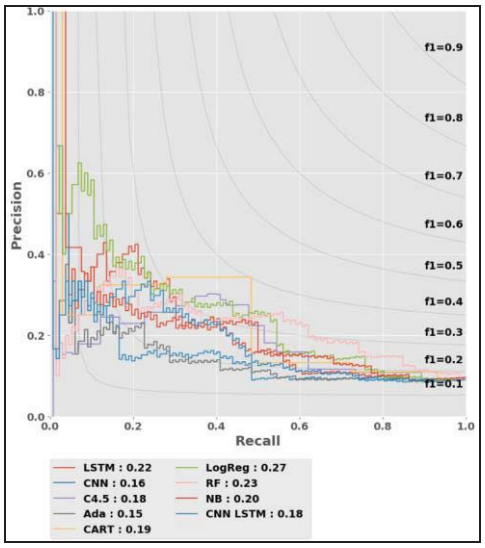

Fig. 5(e). F_1 score (4 hours)

VI. CONCLUSION

The higher the AUROC value or the F_1 score are, the better the model is. Our approach can effectively preprocess the imbalanced dataset and reduce the possible overfitting problem. The performance of different classifiers selected show that the best and the better ones are random forest (RF) and Logistic Regression (LR) respectively when CPR event happened and before that 0 to 4 hours represented by AUROC and F_1 scores. Although RF is the best in 0 hour consistening with that of prior research such as EWS, but it's interesting in that we found that F_1 scores of LR is better than RF in other time periods, 1 to 4 hours, before CPR time as shown in table I. In addition, F_1 scores of LSTM is better than other classifiers and close to that of RF and LR when 1 to 4 hours before the CPR event happened. It's believed that LSTM is suitable for this kind of application regarding the early detection of clinic treatment that the dataset has the property of time sequence. In our observation, 0 hour distancing from the leave time or CPR time is meaningless for it lacks of the meaning and the function of early detection and prediction, and thus the machine or deep learning tachniques except RF is worthy applying adaptively to more applications with the cases having the similar propoerty of this work.

REFERENCES

[1] J. M. Kwon, Y. Lee, Y. Lee, S. Lee, and J. Park, "An Algorithm Based on Deep Learning for Predicting In‐Hospital Cardiac Arrest" Journal of the American Heart Association, vol. 7, no. 13, e008678, 2018.

[2] L. J. Blackhall, "Must we always use CPR" New England Journal Medicine, vol. 317, no. 20, pp. 1281-1285, 1987.

[3] D. L. Atkins, et al., "Part 11: pediatric basic life support and cardiopulmonary resuscitation quality: 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care", Circulation, 132(18 suppl 2), pp. S519-S525, 2015.

[4] H. K. Chang, C. J. Wu, J. H. Liu, W. S. Lim, H. C. Wang, S. I. Chiu, and J. S. Jang, "Early Detecting In-Hospital Cardiac Arrest Based on Machine Learning on Imbalanced Data" in Proceedings of 7th IEEE International Conference on Healthcare Informatiica (ICHI 2019), pp. 57-66, 2019.

[5] T. K. Ho, "Random decision forests" in Document analysis and recognition, in Proceedings of the third international conference, Vol. 1, pp. 278-282, 1995.

[6] D. J. Hand and K. Yu, "Idiot's Bayes---not so stupid after all?" International Statistical Review, vol. 69, no. 3, pp. 385-398, 2001.

[7] H. Zhang,"The optimality of naive Bayes" in Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS '04). AAAI, pp. 562-567, 2004.

[8] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm" in Proceedings of the 13th International Conference on Machine Learning (ICML '96), pp. 148-156, 1996.

[9] R. E. Schapire, "A brief introduction to boosting" in Proceedings of the 16th international joint conference on Artificial intelligence, pp.1401-1406, 1999.

[10] J. R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1993.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory" Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[12] X. Wei, F. Jiang, F. Wei, J. Zhang, W. Liao, and S. Cheng, "An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset" in Proceedings of the computing frontiers conference, pp. 71-78, 2017.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique" Journal of artificial intelligence research, vol 16, pp. 321-357, 2002.

[14] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning" in International Conference on Intelligent Computing, pp. 878-887, 2005.

[15] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning" in Neural Networks, IJCNN 2008, pp. 1322-1328, 2008.

[16] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review" International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, no.04, pp. 687-719, 2009.

[17] N. V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: special issue on learning from imbalanced data sets" SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, vol. 6, no. 1, pp. 1-6, 2004.

[18] B. Avard, H. McKay, N. Slater, P. Lamberth, K. Daveson and I. Mitchell, "Training Manual for The National Early Warning Score and associated Education Programme", 2016. http://www.rcsi.ie/files/facultyofnursingmidwifery/20160811103824_5.2%20NEWS%20Training.pdf

[19] https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2

[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python" Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[21] https://keras.io/

[22] https://www.tensorflow.org/