Detecting Small Bowel Strangulation using Circulating Cell-Free DNA with Machine Learning

Kazutaka Nishiwaki, Takeshi Yamada, Takuma Iwai, Goro Takahashi, Eiji Uchida and Hayato Ohwada

Abstract—**Small bowel obstruction is a common acute abdomen condition which can play a role in death. Small bowel strangulation (SBS) refers to a small bowel obstruction associated with bowel ischemia. Patients with SBS need emergency surgery because the ischemic small bowel can become necrotic in a short time, causing sepsis and death. Nowadays the "gold standard" for diagnosing SBS is via computed tomography (CT) scanned images. However, an easier way to detect SBS is desired among emergency medicine physicians. Thus, we tried to develop a rapid test using circulating cell-free DNA (ccfDNA). ccfDNA is part of the DNA from a cell that died because of apoptosis or necrosis. The size of ccfDNA varies depending on the origination of cell death. If a patient has SBS, long-size ccfDNA would appear in the peripheral blood. We used data including the concentration of ccfDNA in the blood of certain patients as training data to make a support vector machine, a decision tree, and a learned random forest. We evaluated these classifiers using leave-one-out cross-validation. These machine-learning methods performed well. In addition, the decision tree and random forest results indicate that long-size ccfDNA is important for classifying SBS. In this paper, we demonstrate that machine learning can be an alternative method for detecting SBS and that the concentration of ccfDNA, especially long ccfDNA, contributes to detecting SBS.**

Index Terms—**Bioinformatics, circulating cell-free DNA, machine learning, small bowel obstruction.**

I. Introduction

Small bowel obstruction (SBO) is mainly caused by adhesions after abdominal surgery. Small bowel strangulation (SBS) is a subtype of SBO associated with bowel ischemia. Patients with SBS need emergency surgery because the ischemic small bowel can become necrotic in a short time, causing sepsis and death. Early diagnosis can prevent bowel necrosis and improve the prognosis [1], but clinical diagnosis is still very difficult, and its mortality rate is very high, ranging from 20 to 40% [2]. Nowadays the "gold standard" for diagnosing SBS is via contrast enhanced computed tomography (CT) scanned images [2]-[4]. However, an easier way to diagnose SBS is desired because diagnosis of SBS using CT presents difficulties [5]. Moreover, a contrast agent using contrast-enhanced CT may impair renal function in aged patients or patients with dehydration [6]. Therefore, an easier and safer method to diagnose SBS, such as a blood test, is demanded.

It has been reported that circulating cell-free DNA (ccfDNA) derived from apoptotic or necrotic cells exists in peripheral blood. This is generating interest and potential application to clinical medicine for early diagnosis as a biomarker and monitoring of therapy, especially in cancer research [7], [8]. In general, the length of the ccfDNA originating from necrotic cells is greater than that from apoptotic cells [9]. In patients with SBS, the intestinal cells develop necrosis. Thus, long ccfDNA derived from intestinal necrotic cells can be detected in patients with SBS. However, the length that indicates intestinal cell necrosis remains unclear.

In patients with SBS, the amount of long ccfDNA originating in the strangulated bowel is increasing. If identifying bowel strangulation or its absence based on the amount of ccfDNA becomes feasible, patients will not need to be diagnosed by CT equipment.

In this study, we tried to determine the cutoff value for the length of ccfDNA that would strongly indicate bowel necrosis. We used data obtained from a set of patients, including the concentration of ccfDNA in the blood, and applied these data to a support vector machine (SVM) [10], a decision tree [11], and a random forest [12]. SVMs are widely used for classification tasks and have provided good results in many research fields [13], [14]. The decision tree is a simple learning method that has difficulty solving complex problems but can be visualized as a tree model and is easy to understand. The random forest is an ensemble technique containing a multitude of decision trees. This method is also widely used in many research fields and provides good results [15], [16]. In addition, the random forest can indicate which attributes of the data are important for classification.

In this study, we consider machine learning methods that be applied to detecting SBS.

II. Methodology

A. Datasets

We obtained data from 19 patients and 13 out of the 19 patients had SBS. The data includes not only the concentration of each size of ccfDNA in the blood but also attributes such as age, sex, time to treatment, and existence of renal dysfunction. The age range is 41 to 89, and the range of time to treatment is between three hours and one day. We added the time to treatment data as an attribute because ccfDNA has a short half-life. Thus, we set up a hypothesis

Manuscript received February 5, 2017; revised April 6, 2017.

Kazutaka Nishiwaki, and Hayato Ohwada are with Tokyo University of Science, Noda city, Chiba, 278-8510 Japan (e-mail: 7416634@ed.tus.ac.jp, ohwada@rs.tus.ac.jp).

Takeshi Yamada, Takuma Iwai, Goro Takahashi, Kouki Takeda, and Eiji Uchida are with Nippon Medical School, Bunkyo-ku, Tokyo, 113-3603 Japan. (e-mail: y-tak@nms.ac.jp).

doi: 10.18178/ijmlc.2017.7.3.616

that the time to treat may be important for detecting SBS because the amount of both long size and short size ccfDNA would increase in an SBS patient. We also focused on the existence of renal dysfunction because we had formed a hypothesis that ccfDNA would not be excreted from the body, so the concentration of ccfDNA would increase if the kidney were having trouble.

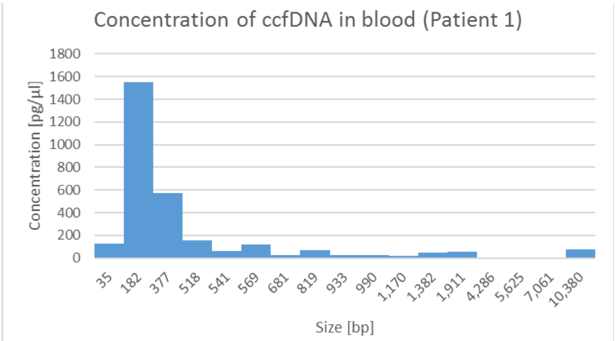


Fig. 1. Concentration of ccfDNA in patient 1.

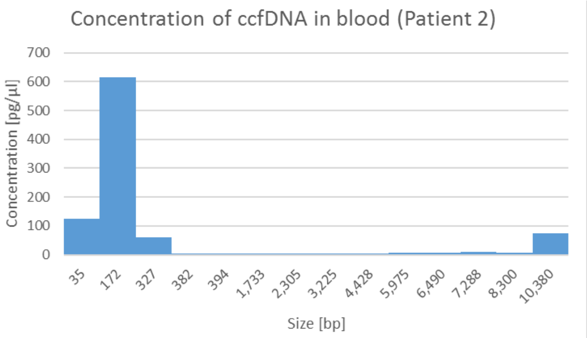
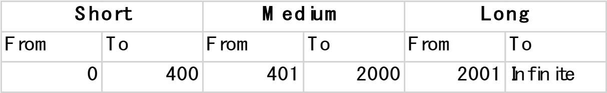
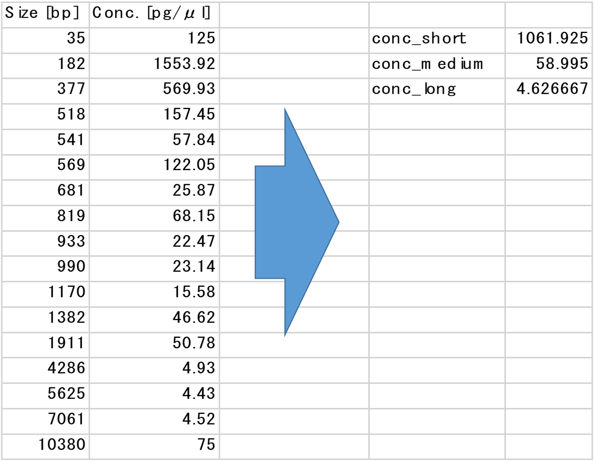


Fig. 2. Concentration of ccfDNA in patient 2.



Short		Medium		Long	
From	To	From	To	From	To
0	400	401	2000	2001	Infinite

Fig. 1. Boundaries for each range.



Size [bp]	Conc. [pg/μl]			
35	125		conc_short	1061.925
182	1553.92		conc_medium	58.995
377	569.93		conc_long	4.626667
518	157.45			
541	57.84			
569	122.05			
681	25.87			
819	68.15			
933	22.47			
990	23.14			
1170	15.58			
1382	46.62			
1911	50.78			
4286	4.93			
5625	4.43			
7061	4.52			
10380	75			

Fig. 2. Example of data conversion.

However, the data representing the ccfDNA concentration in the blood differed depending on the patient. Fig. 1 and Fig. 2, for example, give the concentrations of ccfDNA of each size in the blood of patient 1 and patient 2. The patients have different distributions except for the lower and upper markers (35bp and 10,380bp). These are not good for machine learning because the machine-learning method cannot work with such different dimensions of data. Thus, we had to convert this concentration data into another format. The process for creating a format to make this data accessible to the machine-learning method is as follows.

Determine the threshold for ccfDNA size.

Calculate the mean concentration over a range of sizes partitioned by threshold. We did not take into account the concentration values of 35bp and 10,380bp at this time.

Use that value as a feature for learning.

In this study, we set 400bp and 2000bp as thresholds by intuition based on experiments and observation of all 19 patients' ccfDNA concentration distributions. Fig. 3 presents the boundaries for each range, and Fig. 4 presents an example of this data processing.

B. Machine Learning Methods

In this study, we used SVM, decision tree, and random forest methods to classify patients into those who have SBS and those who do not.

SVM cannot only separate positive and negative examples but can also find a hyperplane that maximizes the margin between positive and negative samples, thus showing a high generalization ability. In addition, SVM can be applied to non-linear problems through the kernel method.

Decision tree is a simple machine-learning method but cannot perform well in complex problems. However, a decision tree can show decision rules (i.e. if-then rules) that are easy for a human to understand. These decision rules are obtained by comparing the information gains of each conceivable rule and then making information gain reach its maximum as each node is used.

A random forest is an ensemble machine-learning method consisting of a multitude of decision trees. Using a multitude of weak learners such as decision trees can reduce the variance so that this ensemble method can treat complex problems. In addition, a random forest can indicate how important each feature is for classifying the positive and negative examples.

Furthermore, we evaluated the threshold for dividing a range of ccfDNA concentrations using decision rules obtained from a decision tree and importance obtained by a random forest. If these methods indicate that short, medium, or long ccfDNA concentrations are important for classifying patients, then these thresholds are correct.

C. Performance Measurement

We used leave-one-out cross-validation to check the classifier performance and calculated the accuracy, precision, recall, and f-measure. These scores can be calculated by the four outcomes of cross-validation: true positive (TP), false positive (FP), false negative (FN), and true negative (TN) (see Table I).

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

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

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

$$F\_measure = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

TABLE. I. CONFUSION MATRIX

		Actual	
		Strangulation	Not
Predicted	Strangulation	TP	FP
	Not	FN	TN

III. RESULTS AND DISCUSSION

We performed experiments using Python 3 in which all of the machine-learning methods were implemented in scikit-learn. The hyper-parameters of these methods were optimized through a grid search. Table II shows the evaluation measures for each method.

TABLE II: PERFORMANCE MEASUREMENT OF EACH METHOD

	SVM	Decision tree	Random forest
Accuracy	0.789	0.895	0.947
Precision	0.800	0.923	0.929
Recall	0.923	0.923	1.000
F measure	0.857	0.923	0.963

The random forest resulted in better scores for all measurements, and in particular a recall of 1.000 was achieved. This shows that random forest could find all SBS cases based on the data for ccfDNA concentrations. In the medical field, a patient has to have surgery if there is even a small possibility that the patient be strangulated, so detecting SBS without omissions is greatly important.

The following sections present details of the prediction results obtained by each method.

A. SVM

Table III shows the confusion matrix obtained for SVM with a radial basis function kernel [17]. The C (gamma) parameter was 10 (0.001).

TABLE III. CONFUSION MATRIX OBTAINED WITH SVM

		Actual	
		Strangulation	Not
Predicted	Strangulation	12	3
	Not	1	3

SVM has been applied widely and performs well. However, it could not classify more correctly than the decision tree and random forest methods. In this study, we optimized only the kernel function, C and the gamma parameters, so better results could be obtained by further optimizing the hyper-parameters.

B. Decision tree

Table IV presents the confusion matrix obtained with the decision tree. We used entropy for the information gain when the decision tree measures the quality of a split.

TABLE IV: CONFUSION MATRIX OBTAINED WITH DECISION TREE

		Actual	
		Strangulation	Not
Predicted	Strangulation	12	1
	Not	1	5

The number of false positives is smaller than that obtained through SVM, so the precision of the decision tree is higher than that of SVM.

Fig. 5 presents the visualized decision tree build using full training data.

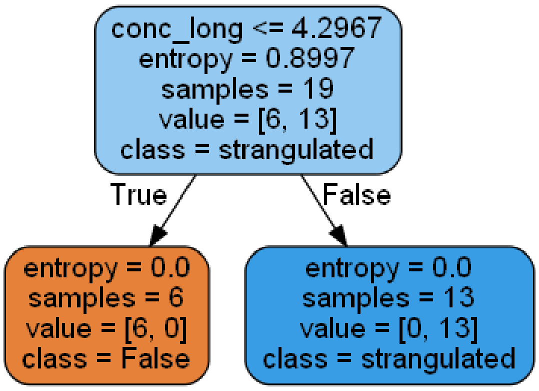


Fig. 3. Visualized decision tree.

Two patients are misclassified in leave-one-out cross-validation. However, we could extract a decision tree that achieved perfect classification by using full training data. That decision rule indicates that the patient must have a strangulated bowel if the average concentration of long-size ccfDNA exceeds 4.2967. This shows that the amount of long ccfDNA in blood is highly important in detecting SBS. In addition, this rule supports the intuition that ccfDNA longer than 2000bp are highly expressed in patients with SBS.

C. Random Forest

Table V shows the confusion matrix obtained with the random forest. As in the case of the decision tree, we used entropy for the information gain. The number of features to consider when looking for the best split and decision trees were the square root of the number of features and 1000. The maximum depth of the decision tree was two.

TABLE V: CONFUSION MATRIX OBTAINED WITH RANDOM FOREST

		Actual	
		Strangulation	Not
Predicted	Strangulation	13	1
	Not	0	5

Random forest achieved zero false negatives and perfect recall. When we focus on detecting SBS, this method is the best choice. Only one non-SBS patient was classified as having SBS. Regarding this patient, the concentration of long-size ccfDNA was not as many as that of an SBS patient but more than others. We consider that this patient was in between the two classes regarding ccfDNA expression.

Table VI shows the importance values of each feature measured by the random forest.

TABLE VI: IMPORTANCE SCORES OF FEATURES

Feature	Importance
Age	0.110
Sex	0.027
Time	0.084
Renal dysfunction	0.043
Conc. short	0.088
Conc. medium	0.283
Conc. long	0.364

The most important feature was the concentration of long-size ccfDNA in the blood, and the critical point was

0.364. This corresponds to the rule obtained through the decision tree. In addition, the concentration of medium-size ccfDNA and age were found to be important for distinguishing SBS and non-SBS patients. However, the existence of renal dysfunction and the time to treat did not affect the detection of SBS through the importance calculated by the random forest in this study.

In this study, we determined that these methods can detect SBS based on the concentration of ccfDNA in the blood. However, the number of patients was extremely small. Machine learning requires a sufficient number of samples to learn. The scikit-learn software indicates that researchers have to prepare more than 50 samples [18], so the number of samples was not sufficient in this study. There is a possibility that these machine-learning methods would work well with a sufficient number of ccfDNA data points.

IV. Conclusion

In this paper, we determined that the SVM, decision tree, and random forest methods all perform well in the task of detecting SBS. In addition, the concentration of ccfDNA in blood was found to contribute to such detection. In an experiment, we calculated the average concentration of each range of ccfDNA size, with the ranges designated as short (less than 400bp), medium (between 401 and 2000bp), and long (more than 2001bp). The experiment results indicated that each machine learning method performed well. The decision tree was able to classify strangulated or normal obstructions by focusing on the average concentration of long ccfDNA. The random forest method determined that the average concentration of long and medium ccfDNA and age contribute to the classification of each state of obstruction. According to these results, applying machine learning would be an alternative way to detect SBS.

Acknowledgement

The authors would like to thank Kouki Takeda for assistance with experiments and data collection.

References

[1] S. Yamagishi et al., "Early Diagnosis of the Strangulated Obstruction," *The Japanese Society of Gastroenterological Surgery* vol. 36, pp. 11-17, 2003.

[2] K. Hayakawa *et al.*, "CT findings of small bowel strangulation: the importance of contrast enhancement," *Emergency Radiology*, vol. 20, pp. 3-9, Jan. 2013.

[3] S. Pothiawala and A. Gogna, "Early diagnosis of bowel obstruction and strangulation by computed tomography in emergency department," *World Journal of Emergency Medicine*, vol. 3, issue 3, pp. 227-231, Sep. 2012.

[4] G. Ohira *et al.*, "Utility of arterial phase of dynamic CT for detection of intestinal ischemia associated with strangulation ileus," *World Journal of Radiology*, vol. 4, issue 11, pp. 450-454, Nov. 2012.

[5] A. Furukawa *et al.*, "Helical CT in the Diagnosis of Small Bowel Obstruction," *Radiographics* vol. 21, issue 2, pp. 341-355, Mar. 2001.

[6] B. B. Aoki *et al*., "Acute kidney injury after contrast-enhanced examination among elderly," *Revista Latino-Americana de Enfermagem*, vol. 22, no. 4, pp. 637-644, Jul.-Aug. 2014.

[7] T. Yamada *et al.*, "Utility of KRAS mutation detection using circulating cell‐free DNA from patients with colorectal cancer," *Cancer Science*, vol. 107, issue 7, pp. 936-943, Jul. 2016.

[8] Y. I. Elshimali *et al*., "The Clinical Utilization of Circulating Cell Free DNA (CCFDNA) in blood of cancer patients," *International Journal of Molecular Sciences*, vol. 14, issue 9, pp. 18925-18958, Sep. 2013.

[9] O. J. Stötzer, J. Lehner, M Braun, and S. Holdenrieder, "Circulating cell free DNA as blood based biomarker in breast cancer," *Molecular Biology*, vol. 4, issue 1, Aug. 2014.

[10] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proc. the 5th Annual Workshop on Computational Learning Theory*, pp. 144-152, Jul. 1992.

[11] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, issue 1, pp. 81-106, Mar. 1986.

[12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, issue 1, pp. 5-32, Oct. 2001.

[13] D. Chen, H. Bourlard, and J. P. Thiran, "Text identification in complex background using SVM," in *Proc. the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. II621-II626, vol. 2.

[14] I. Guyon, J. Weston, and S. Barnhill, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, issue 1, pp. 389-422, Jan. 2002.

[15] R. Díaz-Uriarte, and S. Alvarez De Andrés, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, issue 3, Jan. 2006.

[16] J. Shotton *et al*., "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *Proc. the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1297-1304.

[17] H. Cao, T. Naito, and Y. Ninomiya, "Approximate RBF kernel SVM and its applications in pedestrian classification," in *Proc. The 1st International Workshop on Machine Learning for Vision-Based Motion Analysis*, Marseille, France, Oct. 2008.

[18] scikit-learn developers. Choosing the right estimator (2016). [Online] Available: http://scikit-learn.org/stable/tutorial/machine_learning_map/

Kazutaka Nishiwaki graduated from the Department of Industry Administration, Faculty of Science and Technology, Tokyo University of Science, Chiba, Japan, in 2016.

Currently he is in the Industrial Administration Master's course at the Graduate School of Science and Technology, Tokyo University of Science. His research interests are bioinformatics, machine learning, and data mining.

Takeshi Yamada graduated from the Hamamatsu University School of Medicine, Shizuoka, Japan, in 1992.

He was an assistant professor from 2013 to 2015 at Nippon Medical School. He has been a associate professor in the Department of Gastrointestinal and Hepato-Biliary-Pancreatic Surgery, Nippon Medical School since 2016. His research interests are gastrointestinal surgery, chemotherapy of gastrointestinal cancer, and abdominal emergency medicine.

Takuma Iwai graduated from the Nippon Medical School Medicine, Tokyo, Japan, in 2007.

Currently he is a graduate student of Nippon Medical School since 2014. His research areas are gastrointestinal surgery, and abdominal emergency medicine.

Goro Takahashi graduated from the Nippon Medical School Medicine, Tokyo, Japan, in 2005.

Currently he is a graduate student of Nippon Medical School from 2014. His research areas are gastrointestinal surgery, and abdominal emergency medicine.

Eiji Uchida graduated from the Nippon Medical School, Tokyo, Japan, in 1976.

He has been a professor in the Department of Gastrointestinal and Hepato-Billiary-Pancreatic Surgery, Nippon Medical School since 2007. His research areas are digestive surgery, chemotherapy of pancreatic cancer, and abdominal emergency medicine.

Hayato Ohwada graduated from the Department of Industrial Administration, Faculty of Science and Engineering, Tokyo University of Science, Chiba, Japan, in 1983. He completed the Doctoral course program with degree at the Division of Science and Engineering Industrial Administration, Tokyo University of Science Graduate School in 1988.

He was a research associate from 1988 to 1998, a lecturer from 1999 to 2000, and an associate professor from 2001 to 2004 at Tokyo University of Science. He has been a professor in the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science since 2005. His research areas are artificial intelligence and intelligent informatics, such as Inductive Logic Programming and web mining.