

OPEN

# Increase *Trichomonas vaginalis* detection based on urine routine analysis through a machine learning approach

Hsin-Yao Wang<sup>1,2,3</sup>, Chung-Chih Hung<sup>1,4,5,6</sup>, Chun-Hsien Chen<sup>1,7</sup>, Tzong-Yi Lee<sup>8,9,10,11</sup>, Kai-Yao Huang<sup>10</sup>, Hsiao-Chen Ning<sup>1,4</sup>, Nan-Chang Lai<sup>1</sup>, Ming-Hsiu Tsai<sup>5</sup>, Li-Chuan Lu<sup>12</sup>, Yi-Ju Tseng<sup>1,7,13</sup> & Jang-Jih Lu<sup>1,3,4</sup>

*Trichomonas vaginalis* (*T. vaginalis*) detection remains an unsolved problem in using of automated instruments for urinalysis. The study proposes a machine learning (ML)-based strategy to increase the detection rate of *T. vaginalis* in urine. On the basis of urinalysis data from a teaching hospital during 2009–2013, individuals underwent at least one urinalysis test were included. Logistic regression, support vector machine, and random forest, were used to select specimens with a high risk of *T. vaginalis* infection for confirmation through microscopic examinations. A total of 410,952 and 428,203 specimens from men and women were tested, of which 91 (0.02%) and 517 (0.12%) *T. vaginalis*-positive specimens were reported, respectively. The prediction models of *T. vaginalis* infection attained an area under the receiver operating characteristic curve of more than 0.87 for women and 0.83 for men. The Lift values of the top 5% risky specimens were above eight. While the most risky vigintile was picked out by the models and confirmed by microscopic examination, the incremental cost-effectiveness ratios for *T. vaginalis* detection in men and women were USD\$170.1 and USD\$29.7, respectively. On the basis of urinalysis, the proposed strategy can significantly increase the detection rate of *T. vaginalis* in a cost-effective manner.

Trichomoniasis, which is caused by *Trichomonas vaginalis* (*T. vaginalis*), affects 30.1 million people in the World Health Organization (WHO) Western Pacific Region and 187.0 million people globally, making it the most prevalent nonviral sexually transmitted infection<sup>1</sup>. *T. vaginalis* has been reported to cause several human infections, typically of the urogenital organs, such as vaginitis, urethritis, and prostatitis<sup>2</sup>. The symptoms of *T. vaginalis* infection vary; up to 50% female patients exhibit no symptoms<sup>2</sup>. *T. vaginalis* infection causes some classical urinary tract infection-like symptoms, such as dysuria and urinary frequency and urgency<sup>3</sup>. Two-thirds of infected individuals remain undiagnosed and untreated<sup>4</sup>. Untreated individuals sustain potential infections lasting from months to years<sup>4</sup>. Typically, this infection does not result in serious sequela. However, in recent years, a number of studies reported some virulence factors of *T. vaginalis* that are associated with severe consequences<sup>5–7</sup>. It means that *T. vaginalis* is not always a self-limited infectious disease. Trichomoniasis has been reported to be related to prostate<sup>8–12</sup> and cervical cancer<sup>13–15</sup>, premature birth<sup>16,17</sup>, and infertility<sup>18</sup>. Furthermore, undiagnosed/untreated

<sup>1</sup>Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan. <sup>2</sup>Ph.D. Program in Biomedical Engineering, Chang Gung University, Taoyuan, Taiwan. <sup>3</sup>School of Medicine, Chang Gung University, Taoyuan, Taiwan. <sup>4</sup>Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taoyuan, Taiwan. <sup>5</sup>Graduate Institute of Technological and Vocational Education, National Taipei University of Technology, Taipei, Taiwan. <sup>6</sup>Department of Laboratory Medicine, Taipei Hospital, Ministry of Health and Welfare, New Taipei City, Taiwan. <sup>7</sup>Department of Information Management, Chang Gung University, Taoyuan, Taiwan. <sup>8</sup>Department of Computer Science & Engineering, Yuan Ze University, Taoyuan, Taiwan. <sup>9</sup>Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan, Taiwan. <sup>10</sup>Warshel Institute for Computational Biology, Chinese University of Hong Kong, Shenzhen, China. <sup>11</sup>School of Science and Engineering, Chinese University of Hong Kong, Shenzhen, China. <sup>12</sup>Department of Pathology, National Defense Medical Center, Division of Clinical Pathology, Tri-Service General Hospital, Taipei, Taiwan. <sup>13</sup>Research Center for Emerging Viral Infections, Chang Gung University, Taoyuan, Taiwan. Hsin-Yao Wang and Chung-Chih Hung contributed equally. Correspondence and requests for materials should be addressed to Y.-J.T. (email: [yjtseng@mail.cgu.edu.tw](mailto:yjtseng@mail.cgu.edu.tw)) or J.-J.L. (email: [janglu45@gmail.com](mailto:janglu45@gmail.com))

trichomoniasis is associated with crucial public health concerns. Despite the importance of *T. vaginalis* infection, accurate diagnosis of *T. vaginalis* infection is yet to be standardized.

Trichomoniasis diagnosis can be enhanced by an adequate screening tool<sup>1,17,19</sup>. However, the cost-effectiveness of screening asymptomatic individuals for *T. vaginalis* infection has not been sufficiently addressed<sup>1,17</sup>. *T. vaginalis* can be detected using various test methods, including the microscopic examinations of urine sediments, wet preparation of genital secretions, polymerase chain reaction (PCR), and antigen–antibody rapid screening. Wet preparation of genital secretions or wet mount is the diagnostic method of choice recommended by the Centers for Disease Control and Prevention. The wet mount method includes specimen collection from the vagina or urethra followed by staining and microscopic examination<sup>4</sup>. A good wet mount test largely depends on the adequate collection of vaginal discharge, which should be performed by well-trained medical staff. Consequently, the compliance of wet mount is restricted by limited medical staff. PCR methods<sup>20,21</sup> and antigen–antibody rapid screening<sup>22</sup> could detect *T. vaginalis* with high accuracy. However, the availability and cost-effectiveness of these tests limit their use in routine diagnostic laboratories<sup>23</sup>. By contrast, the microscopic examination of urine remains one of the most commonly used methods for *T. vaginalis* screening<sup>24</sup>, although it is less sensitive than other methods<sup>25</sup>. Microscopic examination of urine is a part of routine urinalysis test, which also tests a number of items, including leukocyte esterase, nitrite, protein, occult blood, red blood cell (RBC) count, white blood cell (WBC) count, epithelial cell count, and sediments in urine specimens. To date, universal microscopic examinations for urine sediments seem to be impractical because of the increasing specimens and limited medical resources<sup>18</sup>. A visual examination of every specimen by medical technologists is an extremely labor-intensive and time-consuming task<sup>19</sup>. Therefore, to examine overwhelming specimens, the routine urinalysis test was automated. Automated instruments can shorten the processing time, reduce the requirement of human resources, and considerably improve efficiency<sup>20–23</sup>. However, all these automated instruments have limitations in detecting *T. vaginalis*<sup>23,24</sup>. Prior to the use of automated instruments, the detection rate of *T. vaginalis* achieved through the visual examination of every urine sediment specimen was approximately 0.1%, based on the data obtained from Chang Gung Memorial Hospital (CGMH), Linkou branch. The detection rate dropped to almost zero, and the role of *T. vaginalis* screening was compromised after the introduction of automated instruments for urine sediment examinations.

Thus, given the fact that all of the routine urinalysis tests have been automated, we intended to improve *T. vaginalis* detection on the basis of other urinalysis test results (i.e., leukocyte esterase, nitrite, protein, occult blood, and RBC, WBC, and epithelial cell counts). We hypothesized that a specific pattern of urinalysis test results for *T. vaginalis*-positive cases versus *T. vaginalis*-negative cases would be noted. To address the pattern recognition problem, we utilized machine learning (ML) algorithms trying to identify the specific pattern of urinalysis test results in *T. vaginalis* infection. ML methods are algorithms that can classify unknown cases by learning the multivariable pattern of training cases<sup>26</sup>. The successful application of ML algorithms in biomedical research or clinical use has been reported and raised considerable attention in recent years. Most of the applications were reported in the field of radiology<sup>27</sup>, dermatology<sup>28</sup>, ophthalmology<sup>29</sup>, oncology<sup>30,31</sup>, and anatomic pathology<sup>32</sup>. Still others utilized ML algorithms in analyzing mass data in the field of microbiology<sup>33–35</sup>. To date, however, only a few applications of ML algorithms for analyzing real-world laboratory data have been published<sup>26,36</sup>. The study would be the first to use the routine urinalysis data to increase *T. vaginalis* detection. In this study, we developed a strategy for optimizing specimen selection for microscopic examinations to facilitate *T. vaginalis* detection. We propose an ML-based strategy for predicting *T. vaginalis*-infected specimens based on the data obtained from automated urinalysis. Through the use of our ML-based strategy, the detection rate of *T. vaginalis* could be increased in a cost-effective manner.

## Results

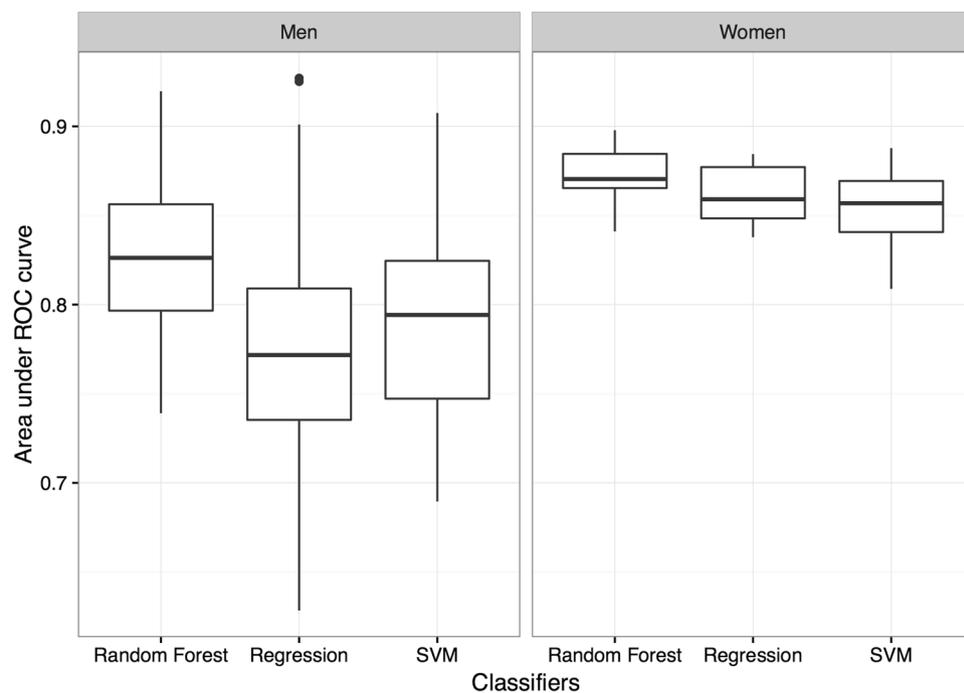
**Patient characteristics and urinalysis results.** A total of 410,952 and 428,203 specimens from men and women were tested, of which 91 (0.02%) and 517 (0.12%) *T. vaginalis*-positive specimens were reported, respectively. Table 1 presents the demographic characteristics and urinalysis results. *T. vaginalis*-positive specimens were more likely to have a higher level of leukocyte esterase and protein, as well as higher WBC and epithelial cell counts, in both women and men, compared with *T. vaginalis*-negative specimens ( $p < 0.001$ ). Women with *T. vaginalis*-positive specimens were younger than those with *T. vaginalis*-negative specimens. By contrast, men with *T. vaginalis*-positive specimens were older than those with *T. vaginalis*-negative specimens. The distributions of continuous and noncontinuous urinalysis test results were showed in Supplementary Figs 1, 2, respectively.

**Model performance.** For *T. vaginalis* detection, the most favorable models for the testing set (50 times, 5-fold cross validation with 10 randomly selected training dataset, Supplementary Fig. 3) were constructed using random forest, with the corresponding area under the receiver operating characteristic (ROC) curve (AUC) values being 0.87 and 0.83 for women and men, respectively ( $p < 0.001$ , Fig. 1). The result shows that the random forest model is a very good model for *T. vaginalis* detection<sup>37,38</sup>. The AUC values derived for multivariable models constructed using random forest, linear regression, and support vector machine (SVM) were significantly higher ( $p < 0.001$ ) than those derived for single-variable models (Supplementary Table 1). Figure 2 shows prediction models' Lift values in each vigintile. The Lift values in the highest vigintile of risk were 8.41 and 8.38 for women and men, respectively, indicating that the positive predictive value in the highest vigintile of risk was more than eight times higher than the average positive predictive value.

**Variable importance.** In the optimal random forest models for *T. vaginalis* detection in women, the most important urinalysis tests are leukocyte esterase, and WBC and epithelial cell counts (Fig. 3). The average mean decrease in Gini of these features was much higher than the other features in urinalysis tests. The features used in the model for men are similar (Fig. 3). The most important features were age, and WBC and epithelial cell counts. Leukocyte esterase was less important in the model for men. Although the RBC count is not significantly

	Women			Men		
	Trichomonas (+)	Trichomonas (-)	P value	Trichomonas (+)	Trichomonas (-)	P value
Patients, n	517	427,686		91	410,861	
Age, mean (SD <sup>a</sup> )	43.0 (15.7)	47.9 (23.0)	<0.001 <sup>c</sup>	63.5 (13.5)	48.5 (24.3)	<0.001 <sup>c</sup>
Leukocyte esterase, median (IQR <sup>b</sup> )	3 (2)	0 (0)	<0.001 <sup>d</sup>	1 (2)	0 (0)	<0.001 <sup>d</sup>
Nitrite, n (%)	35 (6.8)	35987 (8.4)	0.20 <sup>e</sup>	7 (7.7)	19,131 (4.6)	0.26 <sup>e</sup>
Protein, median (IQR)	0 (2)	0 (1)	<0.001 <sup>d</sup>	1 (3)	0 (2)	<0.001 <sup>d</sup>
Occult blood, median (IQR)	1 (3)	0 (2)	<0.001 <sup>d</sup>	0 (2)	0 (2)	0.09 <sup>d</sup>
RBC, mean (SD)	54.4 (128.4)	34.4 (103.0)	<0.001 <sup>c</sup>	37.6 (114.6)	37.7 (112.2)	1.00 <sup>c</sup>
WBC, mean (SD)	129.1 (159.4)	47.2 (113.8)	<0.001 <sup>c</sup>	81.6 (141.4)	26.2 (90.1)	<0.001 <sup>c</sup>
Epithelial cell, mean (SD)	33.1 (30.5)	9.6 (16.8)	<0.001 <sup>c</sup>	8.5 (15.9)	1.6 (5.2)	<0.001 <sup>c</sup>

**Table 1.** Demographic Characteristics and Urinalysis Results of Individuals with *Trichomonas vaginalis*-Positive and *Trichomonas vaginalis*-Negative Specimens. <sup>a</sup>SD: Standard deviation; <sup>b</sup>IQR: Interquartile range; <sup>c</sup>Student t test; <sup>d</sup>Mann–Whitney U test; <sup>e</sup>Pearson chi-squared test.



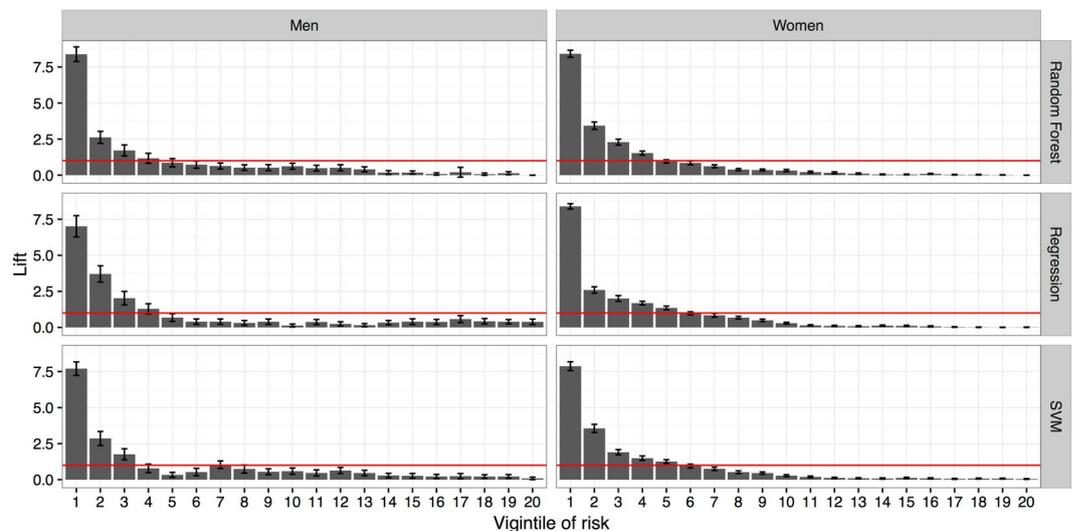
**Figure 1.** Performance of classification models constructed using random forest, linear regression, and SVM classifiers. ROC: Receiver operating characteristic, SVM: Support vector machine.

different between *T. vaginalis*-positive and *T. vaginalis*-negative specimens, the importance of the RBC count in the models were similar with age, contributing to the machine learning models.

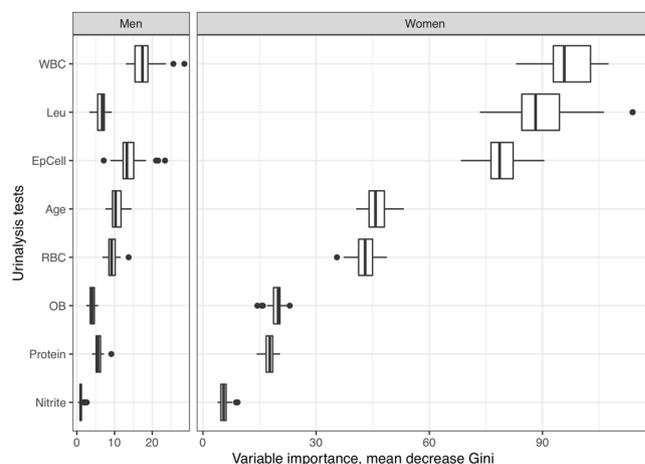
**Cost-effectiveness analysis.** Figure 4 shows plots of the sensitivity versus the percentage of specimens confirmed by microscopic examinations. The sensitivity levels of the classification models constructed using random forest were 83.5% and 74.2% in microscopically examined specimens of women and men, respectively, in the highest quartile of risk. Figure 5 shows plots of the incremental cost-effectiveness ratio (ICERs), the number of specimens microscopically examined divided by the number of positive cases found, versus the sensitivity. The ICERs demonstrated the cost (number of specimens tested) at different levels of effectiveness (number of positive cases found). At a sensitivity threshold of 75%, the lowest ICERs were 197.4 and 1591.4 for women and men, respectively, for the models constructed using random forest.

## Discussion

Urine specimens from individuals at a relatively high risk of *T. vaginalis* infection were successfully identified by the proposed detection model by using data from routine clinical practice. The performance and ICER evaluations revealed that our ML-based strategy can significantly increase the detection rate of *T. vaginalis* in a cost-effective manner.



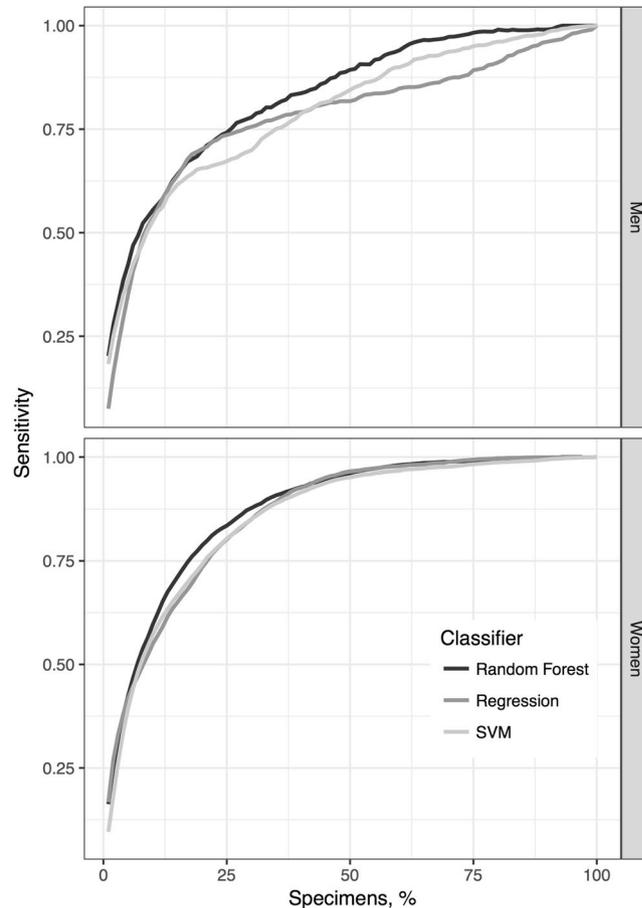
**Figure 2.** Vigintile-wise Lift chart of classification models for men and women. The horizontal lines in the subfigures indicate the average positive predictive value. SVM: Support vector machine.



**Figure 3.** Variable importance of the optimal random forest model, defined by the mean decrease in Gini.

Automation of urine sediment examinations can increase throughput and decrease labor. However, automated instruments could fail to detect some objects including *T. vaginalis*<sup>39,40</sup>. *T. vaginalis* detection methods have received considerable attention because this parasite is associated with various diseases<sup>1,41</sup>. The study variables obtained from automated urinalysis can be used for facilitating *T. vaginalis* detection (Table 1). However, the detection performance associated with using a single variable was not adequately satisfactory (Supplementary Table 1). By contrast, multivariable analysis may result in improved performance, because multivariable analysis provides more information than univariable analysis<sup>26</sup>. The ML models can serve as a sophisticated decision support tool for detecting *T. vaginalis* in urine specimens through training and validation by using a considerably large amount of data.

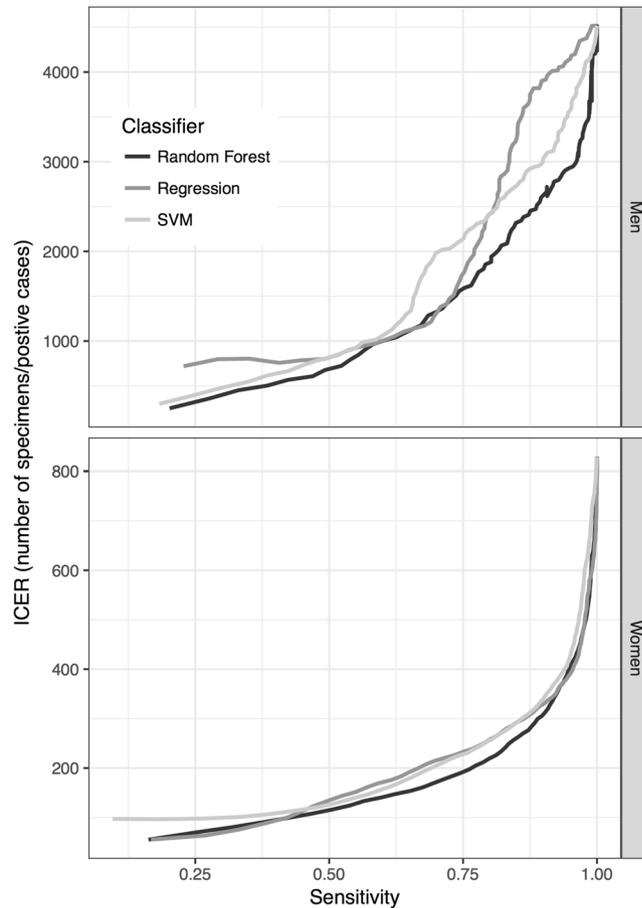
The data used in this study were collected from routine clinical practice. We used all features of urinalysis test to develop and evaluate the ML-based *T. vaginalis* detection strategy and did not exclude the tests which were not significantly different between *T. vaginalis*-positive and *T. vaginalis*-negative specimens because these variables could be important in the full multivariable model<sup>42</sup>. The label of every case was confirmed through microscopic examinations performed by sophisticated medical technologists with annual capacity assessment and College of American Pathologists (CAP) Laboratory Accreditation Program. Although the quality of data was optimized as much as possible, the possibility of false negative cases could not be completely excluded. A confirmatory test for *T. vaginalis* was not performed for several reasons. First, performing nucleic acid-based confirmatory tests for all specimens (i.e., approximately 800,000 cases in 5 years) would have been costly. Moreover, the ML algorithms used in this study are noise-resistant. Therefore, these methods can tolerate falsely labeled cases to some extent<sup>43–45</sup>. Consequently, the data were considered to be qualified for training robust ML models. All ML models achieved an AUC value of more than 0.85 in women (Fig. 1). The performance is sufficient to allow model



**Figure 4.** Curves showing sensitivity achieved by classification models at different percentages of microscopically examined specimens.

application in clinical practice. By contrast, the performance of the ML models in men was lower than that in women. The lower AUC and higher standard deviation may have resulted from much fewer positive cases in men (91 cases) than those in women (517 cases; Table 1). As we know, a robust machine learning model depends largely on data with high fidelity and a sufficient sample size. In this work, although the data were collected over a period of 5 years from a reference hospital, we collected only 91 *T. vaginalis*-positive male cases. The paucity of *T. vaginalis*-positive cases in the male population could be explained by the fact that trichomoniasis is a female-predominant infectious disease. More data are required to construct highly robust models for men.

Urine sediment tests performed in our hospital were not intended for detecting *T. vaginalis*. However, we intended to increase *T. vaginalis* detection by urine sediment screening because it is one of the most frequently requested tests in our hospital (160,000 test requests per year on average). The urine sediment test using microscopic examinations remains a practical method for *T. vaginalis* screening in clinical practice. The advantageous features of this test include easy specimen preparation, immediate results, and low cost<sup>46</sup>. Among the methods capable of detecting *T. vaginalis*, the urine sediment test is ordered more frequently<sup>41</sup>. Therefore, it may be an acceptable tool for *T. vaginalis* screening<sup>25,46</sup>. However, it is extremely labor-intensive. After automation of the urine sediment test in 2015 in CGMH, microscopic examinations have no longer been performed for *T. vaginalis* detection; hence, the *T. vaginalis* detection rate in urine specimens dropped to zero. In the proposed strategy, ML methods were used to score high-risk specimens. We selected only high-risk specimens, identified by ML models, for further microscopic examination. Because overwhelming specimens are received for the urine sediment test (e.g., more than 600 specimens per day in CGMH), a trade-off should be made between the detection rate and the specimens selected for confirmation. If the optimized thresholds of the models obtained from ROC curves are used, approximately 30% of all specimens should be microscopically confirmed. The current manpower in the Department of Laboratory Medicine of CGMH could provide microscopic confirmation for 5% of all specimens following model prediction. The vigintile-wise Lift chart (Fig. 2) revealed that the Lift values were higher than 8 at the first vigintile in both men and women. The classification models provided a much higher probability of detecting *T. vaginalis* than universal microscopic examinations. A review of the top 5% risky specimens demonstrated that a sensitivity of approximately 40% was achieved (Fig. 4; Supplementary Table 2). The results indicate that approximately 40% of the infected cases could be detected through microscopic examinations of one-twentieth of all specimens, thus reducing 95% of the workload. Moreover, under this circumstance of revising top 5% risky specimens, the ICERs were 567 and 99 for men and women, respectively (Fig. 5; Supplementary Table 2). In this study, the ICER was presented as the number of specimens over positive cases to facilitate its utility in different



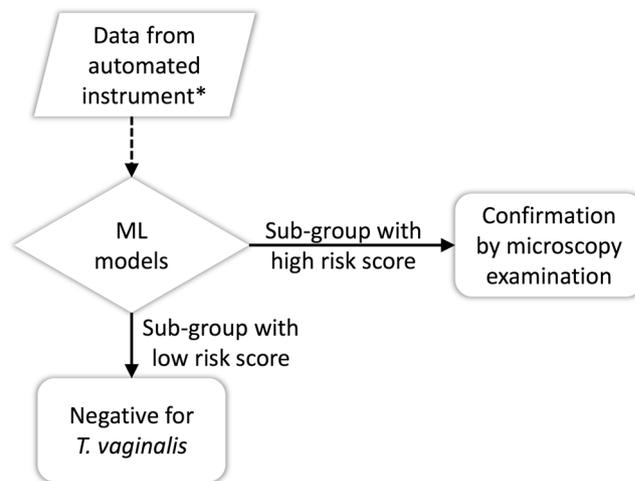
**Figure 5.** Curves showing ICERs at different sensitivity levels. The ICER is defined as the number of specimens microscopically examined divided by the number of positive cases found. ICER: incremental cost-effectiveness ratio.

areas and situations. In CGMH, microscopic examinations of urine sediments cost approximately US\$0.3 per test. Consequently, the costs were determined to be US\$170.1 (i.e.,  $567 \times 0.3$ ) and US\$29.7 (i.e.,  $99 \times 0.3$ ) per positive case in men and women, respectively. The gross domestic product (GDP) per capita of Taiwan was approximately US\$23,000 in 2016<sup>47</sup>. Therefore, *T. vaginalis* detection using the proposed strategy could be considered to be very cost-effective, because the corresponding cost is much lower than the annual GDP per capita, according to the WHO guidelines<sup>48</sup>.

The present study has several limitations. *T. vaginalis* detection was conducted through the microscopic examination of urine sediments. *T. vaginalis* detection in vaginal fluids by using a nucleic acid-based test, such as that reported in the National Health and Nutrition Examination Survey<sup>49</sup>, could have yielded different results. Furthermore, the data in this study were obtained from patients in a reference hospital; the prevalence and incidence might vary with hospitals and regions. The prevalence observed in microscopically examined urine sediments in this study was approximately 0.1%, which is lower than that reported by the WHO or US Centers for Disease Control and Prevention<sup>50</sup>; this discrepancy could be attributed to the aforementioned reasons. The present study successfully demonstrated an ML-based approach for selecting high-risk specimens for further manual detection of *T. vaginalis*. To obtain a highly robust and reliable ML model applicable in clinical practice, nucleic acid-based confirmatory tests may be necessary in the future for assigning a highly accurate label to each urine specimen.

## Methods

**Study population and data.** We performed a retrospective study using laboratory data collected from a 3,383-bed teaching hospital in Taiwan (Chang Gung Memorial Hospital, CGMH) between January 2009 and December 2013. Individuals who underwent at least one urinalysis test [including urine chemistry (URISYS 2400, Roche Diagnostics Corp., Indianapolis, IN, USA) and urine sediment tests through microscopic examination] were included. Cases of *T. vaginalis* infections were identified according to positive *T. vaginalis* test results obtained from microscopic examinations. Individuals with negative test results were considered as controls. If individuals had multiple test results in the study period, all results were included in the analysis. The urinalysis test results consisted of *T. vaginalis*, leukocyte esterase, nitrite, protein, occult blood, RBC count, WBC count, and epithelial cell count. Leukocyte esterase, protein, and occult blood were coded as ordinal integers from 0 to 4 (negative, trace, 1+, 2+, and 3+), 5 (negative, trace, 1+, 2+, 3+, and 4+), and 5 (negative, trace, 1+, 2+, 3+,



**Figure 6.** Flowchart of the machine learning-based strategy for *Trichomonas vaginalis* detection in urine. ML: Machine learning. \*The data included leukocyte esterase, nitrite, protein, occult blood, red blood cell count, white blood cell count, and epithelial cell count.

and 4+), respectively. Positive and negative results for the presence of nitrites and *T. vaginalis* were coded as 1 and 0, respectively. RBC, WBC, and epithelial cell counts were recorded on the basis of the number of cells in a specimen. Sex and age were recorded during the urinalysis test. The Chang Gung Medical Foundation Institutional Review Board approved this study (IRB no. 201601403B0), granting a waiver of patient consent.

**Strategy and model development.** Figure 6 presents a flowchart of the proposed ML-based strategy. Urinalysis data obtained from automated instruments were first analyzed using ML models. The ML models scored the *T. vaginalis* infection risk of each specimen and selected the risky subgroup for confirmation through microscopic examinations. Supplementary Figure 3 shows the framework of ML model development and validation. Patients satisfying the inclusion criteria were randomly assigned to one of five folds. We used a 5-fold cross-validation approach to train (four folds) and test (one fold) the models. To analyze imbalanced data, we randomly selected 10 sets of controls in each round of cross validation, matching the number of cases and age, and generated 10 training datasets by using one set of controls and all cases. Another 5-fold cross-validation process was conducted to tune the classification model in the training step. To account for sex-specific differences, we trained separate models for men and women.

We used logistic regression, SVM with a radial basis function kernel, and random forest to construct predictive models for *T. vaginalis*. Logistic regression measures the relationship between categorical dependent variables and one or more independent variables by using probability scores as the predicted values of the dependent variables<sup>51</sup>. SVM is a data-mining method that constructs a classification model for a binary-class problem. It uses nonlinear mapping to transform the data into a higher dimension. Through an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes are separated by a hyperplane<sup>52</sup>. Random forest is an ensemble classifier proposed by Breiman<sup>43</sup>, comprising many classification trees, the bagging idea, and random selection of features. The frequency of a feature's appearance in classification trees represents the importance of the feature. The models were constructed and tested using R software (version 3.3.2, R Foundation for Statistical Computing, <http://www.r-project.org/>) with the caret package<sup>53</sup>.

**Evaluation methods.** We used the testing set, consisting of one fold of the individuals, to validate the classification models trained by three classifiers with the other four folds of the individuals (Supplementary Fig. 3). The outputs of the model were considered as the risk scores of *T. vaginalis* infection. The AUC and Lift values were used for evaluating model performance. The AUC is a performance measurement for classification problem at various thresholds settings<sup>54</sup>, representing how much the model is capable of distinguishing between classes, interpreted in the ranges of 0.9–1, 0.8–0.9, 0.7–0.8, 0.6–0.7, and 0.5–0.6 as representing excellent, very good, good, sufficient, and fail model, respectively<sup>37,38</sup>. The Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model<sup>54</sup>. For example, suppose a population has an average disease prevalence rate of 5% but a prediction model has identified a high-risk group with a disease rate of 40%. Then, that high-risk group would have a Lift of 8.0 (40% divided by 5%).

**Variable importance.** We evaluated the importance of variables by the mean decrease in Gini<sup>43</sup>. The Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset, computed by summing the probability of an item with a label being chosen times the probability of a mistake in categorizing that item<sup>55</sup>. A split of a node in a tree is made when the Gini impurity criterion for the two descendant nodes is less than that for the parent node. Subsequently, the Gini decreases for each variable over all trees in the forest are summed to determine the variable importance. A higher mean decrease in Gini value represents the greater importance of the variable.

**Cost-effectiveness analysis.** To systematically determine the trade-off between the number of specimens tested and the sensitivity of the classification models, we calculated the sensitivity level at different percentages of specimen tests. Furthermore, we defined the ICER as the number of specimens microscopically examined divided by the number of positive cases found, demonstrating the cost (number of specimens tested) at different levels of effectiveness (number of positive cases found).

**Statistical analysis.** The Student t test, Mann–Whitney U test, and Pearson chi-squared test were used for continuous, ordinal, and categorical data, respectively. Analysis of variance was performed to determine the performance differences among classifiers. All analyses were performed using R software. All statistical tests were two-sided with an  $\alpha$  error level of 0.05.

## Conclusion

The ML-based *T. vaginalis* detection strategy provides a cost-effective means of selecting urine specimens for microscopic examinations, by using multiple urinalysis data obtained from automated instruments. The model can be used in other laboratories that encounter the same problems because of the introduction of automated instruments for urine sediment examinations.

## Data Availability

Data are available from the Ethics Committee of the Chang Gung Memorial Hospital for researchers who meet the criteria for access to confidential data. Requests for the data may be sent to the Chang Gung Medical Foundation Institutional Review Board, Taoyuan City, Taiwan (e-mail: irb1@cgmh.org.tw).

## References

- Meites, E. *et al.* A Review of Evidence-Based Care of Symptomatic Trichomoniasis and Asymptomatic Trichomonas vaginalis Infections. *Clinical Infectious Diseases* **61**, S837–S848, <https://doi.org/10.1093/cid/civ738> (2015).
- Bouchemal, K., Bories, C. & Loiseau, P. M. Strategies for Prevention and Treatment of Trichomonas vaginalis Infections. *Clinical Microbiology Reviews* **30**, 811–825, <https://doi.org/10.1128/cmr.00109-16> (2017).
- Chang, P.-C. *et al.* A pilot study on Trichomonas vaginalis in women with recurrent urinary tract infections. *Biomedical Journal* **39**, 289–294, <https://doi.org/10.1016/j.bj.2015.11.005> (2016).
- Centers for Disease Control and Prevention. *Trichomoniasis - CDC Fact Sheet*, <https://www.cdc.gov/std/trichomonas/stdfact-trichomoniasis.htm> (2017).
- Ryan, C. M., de Miguel, N. & Johnson, P. J. Trichomonas vaginalis: current understanding of host-parasite interactions. *Essays in biochemistry* **51**, 161–175, <https://doi.org/10.1042/bse0510161> (2011).
- Figueroa-Angulo, E. E. *et al.* The effects of environmental factors on the virulence of Trichomonas vaginalis. *Microbes and infection* **14**, 1411–1427, <https://doi.org/10.1016/j.micinf.2012.09.004> (2012).
- Conrad, M. D., Bradic, M., Warring, S. D., Gorman, A. W. & Carlton, J. M. Getting trichy: tools and approaches to interrogating Trichomonas vaginalis in a post-genome world. *Trends in parasitology* **29**, 17–25, <https://doi.org/10.1016/j.pt.2012.10.004> (2013).
- Im, S. J. *et al.* Inflammatory response of a prostate stromal cell line induced by Trichomonas vaginalis. *Parasite immunology* **38**, 218–227, <https://doi.org/10.1111/pim.12308> (2016).
- Iqbal, J., Al-Rashed, J. & Kehinde, E. O. Detection of Trichomonas vaginalis in prostate tissue and serostatus in patients with asymptomatic benign prostatic hyperplasia. *BMC Infectious Diseases* **16**, 506, <https://doi.org/10.1186/s12879-016-1843-1> (2016).
- Kim, J. H. *et al.* Proliferation of Prostate Stromal Cell Induced by Benign Prostatic Hyperplasia Epithelial Cell Stimulated With Trichomonas vaginalis via Crosstalk With Mast Cell. *The Prostate* **76**, 1431–1444, <https://doi.org/10.1002/pros.23227> (2016).
- Shui, I. M. *et al.* Trichomonas vaginalis infection and risk of advanced prostate cancer. *The Prostate* **76**, 620–623, <https://doi.org/10.1002/pros.23153> (2016).
- Zhu, Z. *et al.* Trichomonas vaginalis: a possible foe to prostate cancer. *Medical oncology (Northwood, London, England)* **33**, 115, <https://doi.org/10.1007/s12032-016-0832-y> (2016).
- Lusk, M. J. *et al.* Cervicitis: a prospective observational study of empiric azithromycin treatment in women with cervicitis and non-specific cervicitis. *International journal of STD & AIDS*. <https://doi.org/10.1177/0956462416628319> (2016).
- Parkpinyo, N., Inthasorn, P., Laiwejpithaya, S. & Punnarat, T. Benefits of Cervical Cancer Screening by Liquid-Based Cytology as Part of Routine Antenatal Assessment. *Asian Pacific journal of cancer prevention: APJCP* **17**, 4457–4461 (2016).
- Zemouri, C. *et al.* The Performance of the Vaginal Discharge Syndromic Management in Treating Vaginal and Cervical Infection: A Systematic Review and Meta-Analysis. *PLoS one* **11**, e0163365, <https://doi.org/10.1371/journal.pone.0163365> (2016).
- Moodley, P., Wilkinson, D., Connolly, C., Moodley, J. & Sturm, A. W. Trichomonas vaginalis is associated with pelvic inflammatory disease in women infected with human immunodeficiency virus. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* **34**, 519–522, <https://doi.org/10.1086/338399> (2002).
- Bachmann, L. H. *et al.* Trichomonas vaginalis genital infections: progress and challenges. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* **53**(Suppl 3), S160–172, <https://doi.org/10.1093/cid/cir705> (2011).
- Mielczarek, E. & Blaszkowska, J. Trichomonas vaginalis: pathogenicity and potential role in human reproductive failure. *Infection* **44**, 447–458.
- Thorner, R. M. & Remein, Q. R. *Principles and Procedures in the Evaluation of Screening for Disease*. (Wash., 1961).
- Muzny, C. A., Austin, E. L., Schwebke, J. R., Blackburn, R. J. & Sinsky, R. J. Added Benefit of Nucleic Acid Amplification Testing for the Diagnosis of Trichomonas vaginalis Among Men and Women Attending a Sexually Transmitted Diseases Clinic. *Clinical Infectious Diseases* **59**, 834–841, <https://doi.org/10.1093/cid/ciu446> (2014).
- Harp, D. F. & Chowdhury, I. Trichomoniasis: evaluation to execution. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **157**, 3–9, <https://doi.org/10.1016/j.ejogrb.2011.02.024> (2011).
- Huppert, J. S. *et al.* Use of an Immunochromatographic Assay for Rapid Detection of Trichomonas vaginalis in Vaginal Specimens. *Journal of Clinical Microbiology* **43**, 684–687, <https://doi.org/10.1128/jcm.43.2.684-687.2005> (2005).
- Patil, M., Nagamoti, J. & Metgud, S. Diagnosis of Trichomonas vaginalis from vaginal specimens by wet mount microscopy, in pouch TV culture system, and PCR. *Journal of Global Infectious Diseases* **4**, 22–25, <https://doi.org/10.4103/0974-777x.93756> (2012).
- Tseng, J.-R. *et al.* Clinical Usefulness of 18F-FDG PET/CT for the Detection of Infections of Unknown Origin in Patients Undergoing Maintenance Hemodialysis. *Journal of Nuclear Medicine* **56**, 681–687, <https://doi.org/10.2967/jnumed.114.151696> (2015).
- Lawing, L. F., Hedges, S. R. & Schwebke, J. R. Detection of trichomonos in vaginal and urine specimens from women by culture and PCR. *Journal of clinical microbiology* **38**, 3585–3588 (2000).
- Wang, H.-Y. *et al.* Cancers Screening in an Asymptomatic Population by Using Multiple Tumour Markers. *PLoS one* **11**, e0158285, <https://doi.org/10.1371/journal.pone.0158285> (2016).

27. Kohli, M., Prevedello, L. M., Filice, R. W. & Geis, J. R. Implementing Machine Learning in Radiology Practice and Research. *American Journal of Roentgenology* **208**, 754–760, <https://doi.org/10.2214/AJR.16.17224> (2017).
28. Tschandl, P. *et al.* Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA Dermatology* **155**, 58–65, <https://doi.org/10.1001/jamadermatol.2018.4378> (2019).
29. Ting, D. S. W. *et al.* Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* **318**, 2211–2223, <https://doi.org/10.1001/jama.2017.18152> (2017).
30. Liang, J.-D. *et al.* Recurrence Predictive Models for Patients with Hepatocellular Carcinoma after Radiofrequency Ablation Using Support Vector Machines with Feature Selection Methods. *Computer Methods and Programs in Biomedicine* **117**, 425–434, <https://doi.org/10.1016/j.cmpb.2014.09.001> (2014).
31. Ping, X. O. *et al.* A multiple measurements case-based reasoning method for predicting recurrent status of liver cancer patients. *Computers in Industry* **69**, 12–21, <https://doi.org/10.1016/j.compind.2015.01.007> (2015).
32. Yu, K.-H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* **7**, 12474, <https://doi.org/10.1038/ncomms12474> <https://www.nature.com/articles/ncomms12474#supplementary-information> (2016).
33. Wang, H.-Y. *et al.* Application of a MALDI-TOF analysis platform (ClinProTools) for rapid and preliminary report of MRSA sequence types in Taiwan. *PeerJ* **6**, e5784, <https://doi.org/10.7717/peerj.5784> (2018).
34. Wang, H.-Y. *et al.* A new scheme for strain typing of methicillin-resistant *Staphylococcus aureus* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach. *PLoS one* **13**, e0194289, <https://doi.org/10.1371/journal.pone.0194289> (2018).
35. Wang, H.-Y. *et al.* Rapid Detection of Heterogeneous Vancomycin-Intermediate *Staphylococcus aureus* Based on Matrix-Assisted Laser Desorption Ionization Time-of-Flight: Using a Machine Learning Approach and Unbiased Validation. *Frontiers in Microbiology* **9**, <https://doi.org/10.3389/fmicb.2018.02393> (2018).
36. Lin, W. Y. *et al.* Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. *International journal of medical informatics* **111**, 159–164, <https://doi.org/10.1016/j.ijmedinf.2018.01.002> (2018).
37. Šimundić, A.-M. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* **19**, 203–211 (2009).
38. Momeni, A., Pincus, M., Libien, J. & SpringerLink. *Introduction to Statistical Methods in Pathology*. (Springer, 2018).
39. Shayanfar, N., Tobler, U., von Eckardstein, A. & Bestmann, L. Automated urinalysis: first experiences and a comparison between the Iris iQ200 urine microscopy system, the Sysmex UF-100 flow cytometer and manual microscopic particle counting. *Clinical chemistry and laboratory medicine* **45**, 1251–1256, <https://doi.org/10.1515/cclm.2007.503> (2007).
40. Henneberg, J. R. Comparison between Manual Methods and Automated Analyzer iQ200<sup>®</sup> Iris Diagnostics): A Study for the Optimization of Urinalysis. *International Journal of Laboratory Medicine & Research* **1**, <https://doi.org/10.15344/2455-4006/2015/108> (2015).
41. Tseng, J. R. *et al.* Clinical Usefulness of (1)(8)F-FDG PET/CT for the Detection of Infections of Unknown Origin in Patients Undergoing Maintenance Hemodialysis. *Journal of nuclear medicine: official publication, Society of Nuclear Medicine* **56**, 681–687, <https://doi.org/10.2967/jnumed.114.151696> (2015).
42. Harrell, F. E. Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. (Springer, 2015).
43. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32, <https://doi.org/10.1023/a:1010933404324> (2001).
44. Cruz, J. A. & Wishart, D. S. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics* **2**, 59–77 (2006).
45. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27, <https://doi.org/10.1145/1961189.1961199> (2011).
46. Blake, D. R., Duggan, A. & Joffe, A. Use of spun urine to enhance detection of trichomonas vaginalis in adolescent women. *Archives of Pediatrics & Adolescent Medicine* **153**, 1222–1225, <https://doi.org/10.1001/archpedi.153.12.1222> (1999).
47. Chunghua Institution for Economic Research. *Taiwan's Macroeconomic Forecasts*, <http://www.cier.edu.tw/ct.asp?xItem=12375&CtNode=105&mp=1> (2017).
48. Bertram, M. Y. *et al.* Cost-effectiveness thresholds: pros and cons. *Bull World Health Organ* **94**, 925–930, <https://doi.org/10.2471/BLT.15.164418> (2016).
49. Sutton, M. *et al.* The prevalence of *Trichomonas vaginalis* infection among reproductive-age women in the United States, 2001–2004. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* **45**, 1319–1326, <https://doi.org/10.1086/522532> (2007).
50. Centers for Disease Control and Prevention. *If selected for this survey, your participation is important*, <https://www.cdc.gov/nchs/nhanes/participant.htm> (2017).
51. Hastie, T., Friedman, J. & Tibshirani, R. *The elements of statistical learning*. Vol. 2 (Springer 2009).
52. Cortes, C. & Vapnik, V. Support-Vector Networks. *Machine Learning* **20**, 273–297, <https://doi.org/10.1023/a:1022627411411> (1995).
53. Kuhn, M. Building Predictive Models in R Using the caret Package. *2008* **28**, 26, <https://doi.org/10.18637/jss.v028.i05> (2008).
54. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. *Data Mining: Practical machine learning tools and techniques*. (Morgan Kaufmann 2016).
55. Breiman, L. *Classification and regression trees*. (Routledge 2017).

## Acknowledgements

We acknowledge the members of the Department of Laboratory Medicine of CGMH for their contributions to this study. This manuscript was edited by Wallace Academic Editing. This work was supported by Chang Gung Memorial Hospital [CMRPD3G0111 to Y.T., J.L., C.H. and H.W., CMRPD3I0011 and CORPG3H0441 to Y.T., J.L. and H.W.] and the Ministry of Science and Technology, Taiwan [MOST107-2636-E-182-001 and MOST108-2636-E-182-001 to Y.T.]. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. In addition, the authors report no potential conflicts of interest relevant to this article.

## Author Contributions

H.W. and Y.T. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. H.W. and Y.T. analyzed/interpreted the data, performed experiments, and designed the study and wrote the paper. C.H., C.C., T.L., K.H., H.N., N.L., M.T. and L.L. provided technical, or material support. J.L. and Y.T. obtained finding and supervised the study.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-47361-8>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019