Nuclear Medicine and Molecular Imaging (2019) 53:125–135
https://doi.org/10.1007/s13139-019-00574-1

ORIGINAL ARTICLE

ISSN (print) 1869-3482
ISSN (online) 1869-3474

Development of Predictive Models in Patients with Epiphora Using Lacrimal Scintigraphy and Machine Learning

Yong-Jin Park[1] · Ji Hoon Bae[1] · Mu Heon Shin[1] · Seung Hyup Hyun[1] · Young Seok Cho[1] · Yearn Seong Choe[1] · Joon Young Choi[1] · Kyung-Han Lee[1] · Byung-Tae Kim[1] · Seung Hwan Moon[1]

Received: 12 June 2018 / Revised: 19 September 2018 / Accepted: 7 January 2019 / Published online: 7 February 2019

Abstract

Purpose We developed predictive models using different programming languages and different computing platforms for machine learning (ML) and deep learning (DL) that classify clinical diagnoses in patients with epiphora. We evaluated the diagnostic performance of these models.

Methods Between January 2016 and September 2017, 250 patients with epiphora who underwent dacryocystography (DCG) and lacrimal scintigraphy (LS) were included in the study. We developed five different predictive models using ML tools, Python-based TensorFlow, R, and Microsoft Azure Machine Learning Studio (MAMLS). A total of 27 clinical characteristics and parameters including variables related to epiphora (VE) and variables related to dacryocystography (VDCG) were used as input data. Apart from this, we developed two predictive convolutional neural network (CNN) models for diagnosing LS images. We conducted this study using supervised learning.

Results Among 500 eyes of 250 patients, 59 eyes had anatomical obstruction, 338 eyes had functional obstruction, and the remaining 103 eyes were normal. For the data set that excluded VE and VDCG, the test accuracies in Python-based TensorFlow, R, multiclass logistic regression in MAMLS, multiclass neural network in MAMLS, and nuclear medicine physician were 81.70%, 80.60%, 81.70%, 73.10%, and 80.60%, respectively. The test accuracies of CNN models in three-class classification diagnosis and binary classification diagnosis were 72.00% and 77.42%, respectively.

Conclusions ML-based predictive models using different programming languages and different computing platforms were useful for classifying clinical diagnoses in patients with epiphora and were similar to a clinician's diagnostic ability.

Keywords Epiphora · Dacryocystography · Lacrimal scintigraphy · Machine learning · Deep learning · Convolutional neural network

Introduction

Artificial intelligence (AI) has the potential to revolutionize healthcare. As AI technologies such as machine learning (ML) and deep learning (DL) evolve, there are a growing number of studies that use them in the medical field. ML, a subset of AI, can learn by automatically detecting patterns in training data and then enabling decisionmaking by uncovering patterns when new data is put in. DL, a part of ML, is a special type of artificial neural network that resembles the multilayered human cognition system. These techniques are currently of great interest for utilization with big medical data [1–5].

Nuclear medicine imaging tests can be a good target for the application of AI, and it is expected that improved diagnostic performance and convenience can be obtained through AI techniques. However, there are few studies on AI in the field of nuclear medicine for various reasons. One of them is that it is not easy for general medical practitioners, who do not have a high level of expertise in computer programming, to understand AI techniques and thus to design the system using it. However, with the recent availability of various programming languages and computing platforms for ML and DL, such as

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13139-019-00574-1) contains supplementary material, which is available to authorized users.

✉ Seung Hwan Moon
seunghwan.moons.moon@samsung.com

[1] Departments of Nuclear Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, 50, Irwon-dong, Gangnam-gu, Seoul 135-710, South Korea

Python-based TensorFlow, R, and Microsoft Azure Machine Learning Studio (MAMLS), attempts to use AI in the field of nuclear medicine and other medical fields are becoming increasingly easier. Nonetheless, studies using ML and DL in nuclear medicine tests are rarely reported [6–8]. Above all, there is a lack of reporting on which programming languages or platforms are more accurate and what kinds of advantages or disadvantages are present for each programming language or platform.

Watering eyes, known as epiphora, can be caused by a variety of problems with the lacrimal drainage system. In many situations, the assessment of epiphora in patients whose lacrimal systems are anatomically patent remains challenging [9]. The term "functional obstruction" describes patients with epiphora where the tear duct is anatomically patent and tear drainage is delayed [10]. Lacrimal scintigraphy (LS) is a reliable method of studying the lacrimal drainage system and allows for a more physiological assessment of tear flow, and it is one of the tests that measure the delayed tear drainage. It has traditionally been considered second-line to DCG for investigating epiphora with patency to syringing, and it is used to design more effective procedures for the management of the patients [10, 11]. Furthermore, quantitative values of LS enable the accurate measurement of lacrimal clearance from the conjunctival fornices and allow the application of ML and DL in the diagnosis of functional and nonfunctional epiphora. We expected that LS is a suitable target for the application of ML and DL.

We conducted this study to develop predictive models in patients with epiphora using different programming languages and different computing platforms for ML and DL. We performed supervised learnings from labeled training data sets in Python-based TensorFlow, R, and MAMLS, and we obtained training accuracies and test accuracies. We also confirmed performance of the predictive models and compared them with a nuclear medicine physician. In addition, test accuracy was measured by supervised learning through convolutional neural network (CNN) using LS images only.

Materials and Methods

Subject Selection

We retrospectively reviewed data from 266 patients with epiphora who underwent DCG between January 2016 and September 2017. Among these patients, 16 patients without LS were excluded from this study. A total of 250 patients were enrolled in this study. We used the characteristics of each eye of 250 patients as input variables. This study protocol was reviewed and approved by the ethics committee at our institution.

Lacrimal Scintigraphy

In our institution, LS was performed on patients in the supine position. A dynamic image was obtained for 30 min after instilling two drops of radioactive tracer (1.0 mCi of technetium-99m sodium pertechnetate) into the lateral canthus of the conjunctiva. After maintaining a sitting posture for 10 min, a delayed image was obtained with the patient in the supine position for 1 min. Image acquisition was performed using dual-headed gamma cameras (e.cam, Symbia E, Symbia T16, Siemens, Erlangen, Germany) with a 256 × 256 matrix and a low-energy, high-resolution collimator.

Image Analysis in Lacrimal Scintigraphy

Quantitative analyses were performed on 1-min images and 30-min images by drawing two regions of interest (ROIs) such as the presac to sac area and the postsac area in an analysis software package (Xeleris Functional Imaging Workstation 2.1517, GE Healthcare, USA), and then tabulating counts of radionuclide activity. Visual changes in LS in 30-min and 40-min delayed images were identified.

Five Different Input Variables and an Output Variable

Twenty-seven input variables were used. These included the following: age, sex, hypertension, diabetes mellitus, autoimmune disease, thyroid disease, rhinitis, history of radiation therapy, history of cancer, dry eye, glaucoma, cataract, visual acuity, noncontact intraocular pressure, abnormal tear meniscus, facial palsy, ptosis, treatment history before evaluation, four variables of radioactivity count ratio (presac to sac vs. postsac in 0 min, presac to sac vs. postsac in 30 min, presac to sac in 30 min vs. presac to sac in 0 min, postsac in 30 min vs. postsac in 0 min), and visual changes in LS in 30-min and 40-min delayed images. In addition, two variables related with epiphora (VE) such as the presence of epiphora and the duration of epiphora and two variables related with DCG (VDCG) such as anatomical obstruction or stenosis in DCG for each eye were used.

We classified different sets of input variables depending on whether VE or VDCG are included for four of the five predictive models. We designated model 1 which includes both VE and VDCG and model 4 which excludes both VE and VDCG as input variables. In addition, we designated model 2 which excludes VDCG and model 3 which excludes VE as input variables. When VDCG is included, three-class classification is used. However, if VDCG is not included, binary classification is used because anatomical obstruction is excluded in the output variable. We made a model 5 with input variables obtained through multivariate analysis to control confounding effect and compared it with model 1 containing the included all input variables. In the model 5, three-class

classification is used. This study was a supervised ML study. A clinical diagnosis such as anatomical obstruction, functional obstruction, or normal for each eye was defined as an output variable. Process scheme according to five different input variables is presented in Fig. 1.

Training Data Sets and Test Data Sets

To maximize the information extracted from the training data set, 80% of 500 eyes belonging to 250 patients were used as the training data set. The trained classification models were validated on the remaining 20% of the eyes. We randomly split all of the data into a training data set and test data set using the RAND function in Microsoft Excel 2010. Depending on whether VE and VDCG were included or excluded, four different input variables and output variables were used from model 1 to model 4. In the model 1, model 3, model 5, and three-class classification in CNN, the number of training data sets was 400, the number of test data sets was 100, and three-class classification was used. In the model 2, model 4, and binary classification in CNN, the number of training data sets was 346, the number of test data sets was 93, and binary classification that excluded the data set of the anatomical obstruction was used.

Univariate Analysis and Multivariate Analysis

We conducted univariate analysis and multivariate analysis to control a confounding effect between variables using multinomial logistic regression in R version 3.5.1. For all 27 variables, we found the variables whose p value is less than 0.05 in univariate analysis, and then, multivariate analysis was conducted with these variables. After multivariate analysis, variables whose p value is less than 0.05 were included in model 5.

Python-Based TensorFlow

TensorFlow is an open-source software library for dataflow programming across a range of jobs. It is a representative math library, and it is also used for ML applications such as neural networks. It is used for both research and production at Google.

In this study, the Python-based TensorFlow development environment consists of Python 3.5.4, TensorFlow 1.8.0, Anaconda 4.2.0 64 bit, and Pycharm community edition 2017.2.4 on Windows 10 64 bit. In TensorFlow, the hypothesis function was

$$H(X) = XW + b = Y,$$

where X is an input matrix, W is a weight matrix, b is a bias matrix, and Y is an output matrix. Multinomial logistic regression, also known as softmax regression, was used as the activation function. We used one hot encoding that encoded categorical integer features. A cross-entropy cost function was used as the cost function, and a gradient descent optimizer was used to minimize the cost function iteratively. In order for gradient descent to work, we set the learning rate to 0.001 and the number of iterations to 300,000.

R

R is one of an open source programming language for statistical computing and graphics. We performed ML using R to compare results in Python-based TensorFlow with other programming languages. We implemented multinomial logistic regression known as softmax regression for each model using

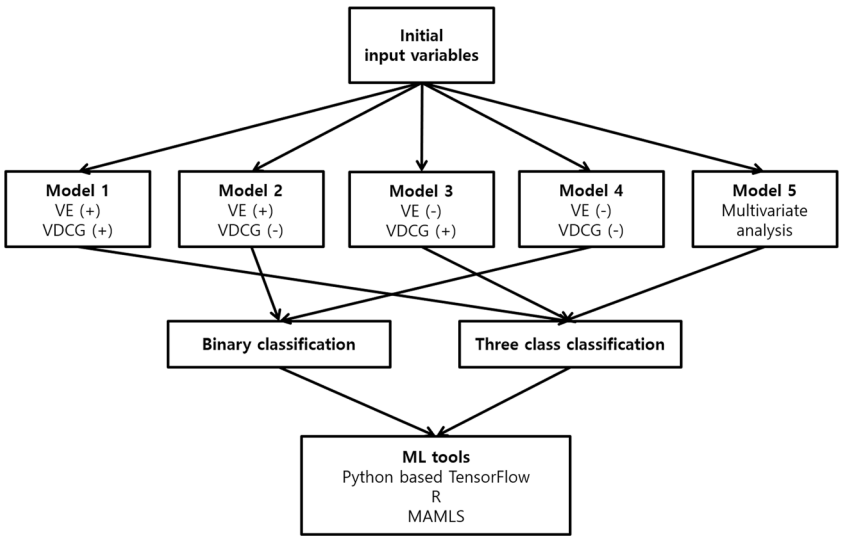

Fig. 1 Process scheme according to five different input variables in five different predictive models. *VE* variables related to epiphora, *VDCG* variables related to dacryocystography, *ML* machine learning, *MAMLS* Microsoft Azure Machine Learning Studio

"nnet" package and RStudio 1.1.453 in R version 3.5.1. Using multinomial logistic regression of nnet package, there was no need to specify hypothesis function and cost function. The package calculated automatically without entering numerical values for learning rate and number of iterations, and then produced results about each predictive model. We obtained variable importance by using the "caret" package after a multinomial logistic regression.

Microsoft Azure Machine Learning Studio

The MAMLS is a cloud-based computing platform that is approachable through a web-based interface. A variety of ML algorithms are obtainable through classification models developed by Microsoft.

In a multiclass logistic regression, we set the module parameters, including the optimization tolerance to 10^{-7}, L1 regularization weight to 1, L2 regularization weight to 1, and memory size for a limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm to 20. An example of multiclass logistic regression in the MAMLS platform is shown in Fig. 2. In a multiclass neural network, we have chosen several options to create neural network. We selected a single parameter in create trainer mode to perform supervised learning without the hyperparameter tuning process. We chose a fully connected case as the hidden layer specification. The number of nodes in input layer was the number of features in training data, and the number of nodes in hidden layer was 100. For three-class classification, the number of nodes in output layer was three, and for binary classification, the number of nodes in output layer was two. Input layer, hidden layer, and output layer were all fully connected. With a learning rate of 0.001 and the number of iterations was 300,000, we made a learning rate and a number of iterations similar to Python-based TensorFlow. An initial learning weight was 0.001, and type of normalizer was min-max normalizer to use feature normalization. As in Python-based TensorFlow, we have chosen cross entropy as cost function.

Convolutional Neural Network

We developed a CNN architecture with input data sets for each eye that were labeled with classification. The CNN development environment consisted of Python 3.5.4, TensorFlow 1.8.0, Keras 2.2.0, and Numpy 1.14.5 on Windows 10 64 bit. The images for each eye were resized to 64 × 64 and were put into CNN network as input image. The CNN architecture consisted of two convolutional layers and two max pooling layers. Kernel size was 3 × 3 in convolutional layers, and pool size was 2 × 2 in max pooling layer. Batch normalization after the convolutional layer was implemented, and the relu function was used as activation function. Two fully connected layers with 128 and 64 nodes and 50% dropout were used to calculate classification scores. An adam optimizer was used as the gradient descent optimization algorithm, and a cross entropy cost function was used as the cost function. Three-class

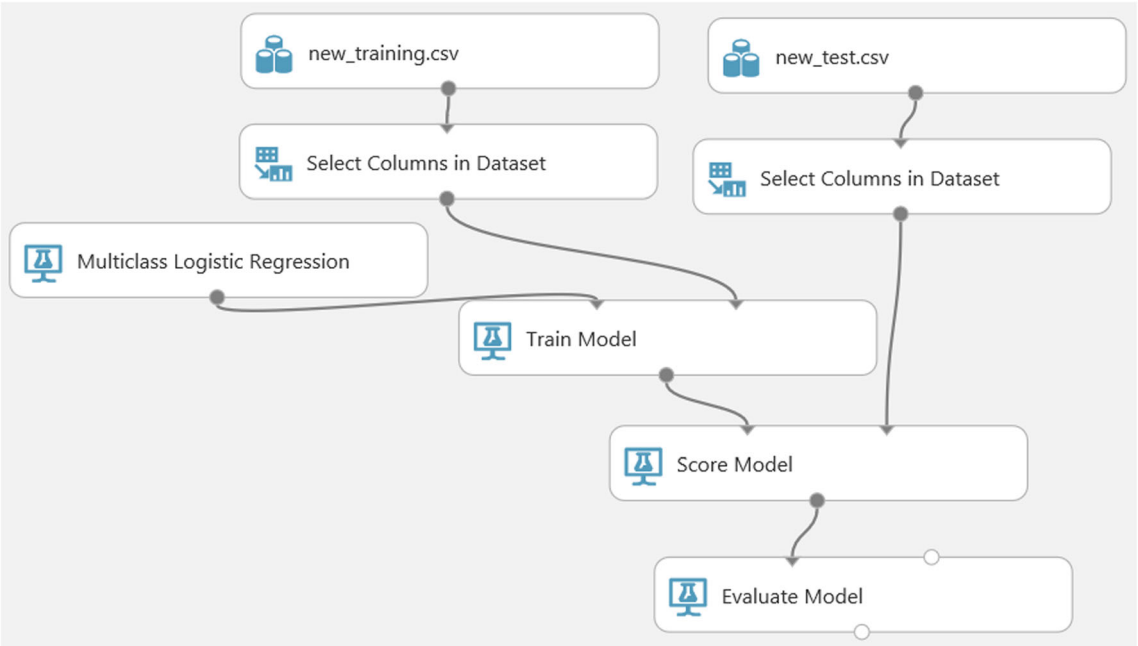

Fig. 2 An example of multiclass logistic regression using MAMLS with training data set and test data set. *MAMLS* Microsoft Azure Machine Learning Studio

classification was implemented to classify the anatomical obstruction, functional obstruction, and normal, and binary classification was also implemented to classify the functional obstruction and normal. The model has done 100 epochs for supervised learning. The CNN architecture is summarized in Fig. 3.

Performance Evaluation of Predictive Models

We obtained accuracy, precision, recall, and F1 score for each predictive model to assess the performance using Python 3.5.4, Pandas 0.23.1, and Scikit-learn 0.19.1. Accuracy is defined as the ratio of the number of correctly classified subjects to the total number of classified subjects. Precision is defined as the number of true positives divided by the number of true positives and false positives. Recall is defined as the number of true positives divided by the number of true positives and false negatives. In general, F1 score is defined as the harmonic mean of precision and recall. We obtained receiver operating characteristic (ROC) curve and area under the curve (AUC) for each classification except for the ones that were perfectly correct in each predictive model using R version 3.5.1 and "ROCR" package.

Statistical analyses were performed using the SPSS software (ver. 20; IBM Corp., Armonk, NY, USA). We performed a McNemar test to determine whether predictive models in different languages or other platforms are statistically equivalent. p values < 0.05 were considered statistically significant.

Analytical Approach to LS of Nuclear Medicine Physician

LS was assessed visually by an experienced nuclear medicine physician who was versed in model 4. The experienced nuclear medicine physician classified whether it was functional obstruction or normal.

Results

Characteristics of Patients and Each Eye

The characteristics of patients and each eye are presented in Table 1. Among 250 patients, 86 (34.4%) were male, 164 (65.6%) were female, and the mean age was 61 ± 13 years. Among each of the 500 eyes of the 250 patients, 59 eyes (11.8%) were clinically diagnosed with anatomical obstruction, 338 eyes (67.6%) had functional obstruction, and 103 eyes (20.6%) were normal. Three hundred ninety-eight eyes (79.6%) had epiphora with a mean duration of 32 ± 49 months.

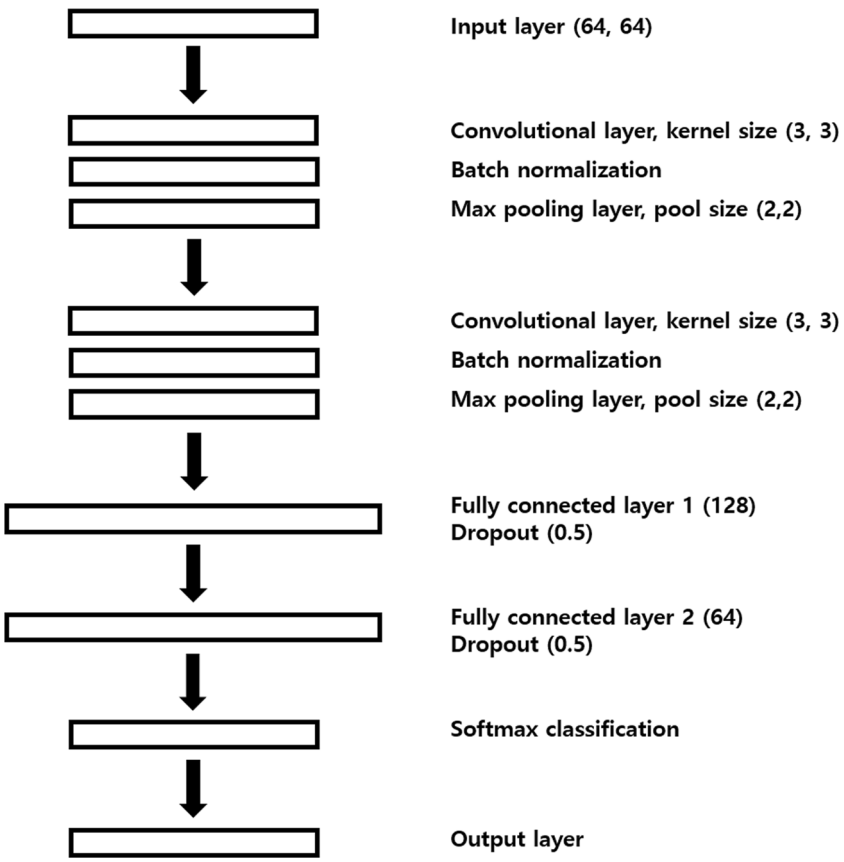

Fig. 3 A summary of CNN architecture. *CNN* convolutional neural network

Table 1 Characteristics of patients and each eye

	Total ($N = 250$)		Total ($N = 500$)
Age (years)		Epiphora	398 (79.6%)
Range	15–87	Duration of epiphora (months)	
Mean ± SD	61 ± 13	Mean ± SD	32 ± 49
Sex		Anatomical obstruction in DCG	61 (12.2%)
M	86 (34.4%)	Stenosis in DCG	196 (39.2%)
F	164 (65.6%)	Dry eye	33 (6.6%)
Hypertension	63 (25.2%)	Glaucoma	14 (2.8%)
Diabetes mellitus	39 (15.6%)	Cataract	169 (33.8%)
Thyroid disease	15 (6.0%)	Visual acuity	
Rhinitis	19 (7.6%)	Mean ± SD	0.7 ± 0.3
History of radiation therapy	7 (2.8%)	Noncontact IOP (mmHg)	
History of cancer	37 (14.8%)	Mean ± SD	16.1 ± 3.9
		Abnormal tear meniscus	367 (73.4%)
		Facial palsy	9 (1.8%)
		Ptosis	9 (1.8%)
		Treatment history before evaluation	153 (30.6%)
		Clinical diagnosis	
		Anatomical obstruction	59 (11.8%)
		Functional obstruction	338 (67.6%)
		Normal	103 (20.6%)

DCG dacryocystography, *IOP* intraocular pressure

In DCG, the lacrimal drainage systems of 61 eyes (12.2%) were obstructed anatomically, and those of 196 eyes (39.2%) were narrowed.

There were inconsistencies between anatomical obstruction in DCG and anatomical obstruction as clinical diagnosis in this study. In the first case, DCG reported an anatomical obstruction to the proximal portion of the nasolacrimal duct. However, there was no epiphora and clinically diagnosed as normal eye. In the second case, DCG reported probably an anatomical obstruction of nasolacrimal duct. This eye had epiphora; however, irrigation reported that the nasolacrimal duct was not obstructed completely and clinically diagnosed as functional obstruction. Except for these two cases, the remaining 59 cases were matched with anatomical obstruction in DCG and anatomical obstruction as clinical diagnosis.

Univariate Analysis and Multivariate Analysis for Model 5 and Variable Importance

Of the 27 variables, we found the 11 variables whose p value was less than 0.05 through univariate analysis. These included the following: presence of epiphora ($p < 0.001$), duration of epiphora ($p < 0.001$), stenosis in DCG ($p < 0.001$), visual acuity ($p = 0.023$), abnormal tear meniscus ($p < 0.001$), history of cancer ($p = 0.003$), treatment history before evaluation ($p < 0.001$), visual change in LS in 30-min vs. 40-min delayed image ($p < 0.001$), count ratio of postsac in 30 min vs. postsac in 0 min ($p = 0.042$), count ratio of presac to sac vs. postsac in 0 min ($p = 0.005$), and count ratio of presac to sac vs. post sac in 30 min ($p = 0.002$). We conducted a multivariate analysis with the 11 variables obtained from univariate analysis and found five variables whose p value is less than 0.05. These included the following: presence of epiphora ($p = 0.002$), stenosis in DCG ($p < 0.001$), history of cancer ($p = 0.006$), visual change in LS in 30-min vs. 40-min delayed images ($p = 0.003$), and count ratio of postsac in 30 min vs. postsac in 0 min ($p = 0.045$). We developed the model 5 with only these five independent variables and compared them to the model 1 with all independent variables.

Variable importance of the 27 variables were obtained using a multinomial logistic regression in R. High variable importance was identified in the following five variables: anatomical obstruction in DCG (1475.94), presence of epiphora (512.24), count ratio of presac to sac in 30 min vs. presac to sac in 0 min (348.89), stenosis in DCG (232.54), and sex (220.31). These results are summarized in Fig. 4.

Comparison of Performance Evaluation of Five Different Predictive Models Using ML Tools

There was no statistically significant difference between the test accuracy of the predictive models using different ML tools and of nuclear medicine physician except in models 1 and 4. In model 1, the test accuracies of Python-based TensorFlow,

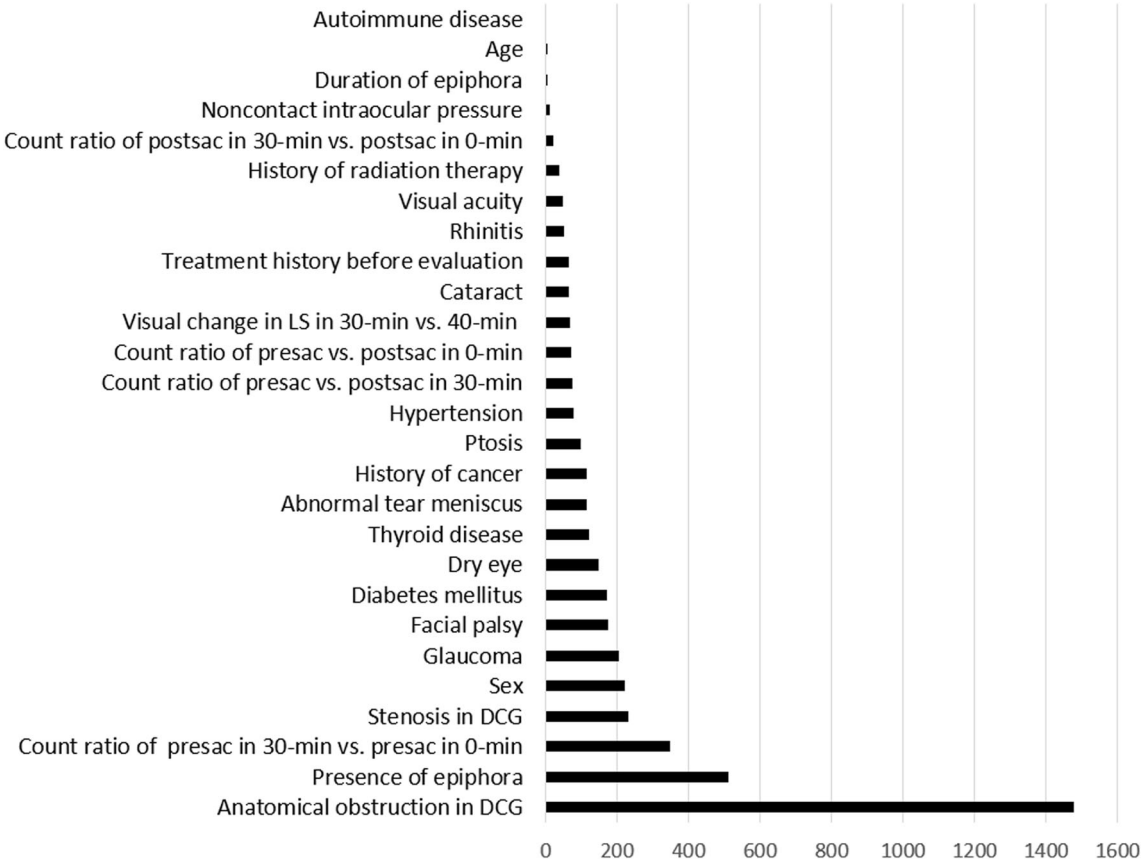

Fig. 4 Variable importance for all 27 independent variables after a multinomial logistic regression in R. *DCG* dacryocystography

R, multiclass logistic regression in MAMLS, and multiclass neural network in MAMLS were 93.00%, 97.00%, 100.00%, and 99.00%, respectively. The predictive models using multiclass logistic regression and multiclass neural network in MAMLS showed higher accuracy compared than Python-based TensorFlow ($p = 0.016$, $p = 0.031$, respectively). There was no significant difference between R and MAMLS, and Python-based TensorFlow and R. In model 4, the test accuracies of Python-based TensorFlow, R, multiclass logistic regression in MAMLS, multiclass neural network in MAMLS, and nuclear medicine physician were 81.70%, 80.60%, 81.70%, 73.10%, and 80.60%, respectively. Compared to the predictive model using multiclass neural network in MAMLS, the others showed higher accuracy ($p = 0.008$, $p = 0.013$, $p = 0.001$, $p < 0.001$, respectively). The accuracy, precision, recall, and F1 score of predictive models from model 1 and model 5 are summarized in Table 2, and test accuracy, precision, recall, and F1 score of the ML tools and nuclear medicine physician in model 4 are summarized in Table 3. Learning curves of the training costs, training accuracies, test costs, and test accuracies from model 1 to model 4 in Python-based TensorFlow during iteration steps are plotted in Fig. 5, and learning curves of the training costs, training accuracies, test costs, and test accuracies in model 1 and model 5 in Python-based TensorFlow during iteration steps are plotted in Supplementary Fig. 1. ROC curves and AUCs for each classification except for the ones that were perfectly correct from model 1 to model 5 between ML tools are summarized in Supplementary Fig. 2 and Supplementary Table 1. ROC curves for functional obstruction in model 1 and model 5 between ML tools are summarized in Supplementary Fig. 3.

Convolutional Neural Network

Test accuracy in the three-class classification of CNN, which classifies the anatomical obstruction, functional obstruction, and normal, was 72.00%. In addition, test accuracy in binary classification of CNN, which classifies functional construction and normal, was 77.42%. The binary classification represented a relatively higher test accuracy than three-class classification.

Discussion

We created predictive models for detecting undiagnosed epiphora using given input variables or input images with different supervised ML tools and CNN, and we confirmed

Table 2 Performance of different predictive models in Python-based TensorFlow, R, and MAMLS

	Model 1 VE (+) VDCG (+)	Model 2 VE (+) VDCG (−)	Model 3 VE (−) VDCG (+)	Model 4 VE (−) VDCG (−)	Model 5 Multivariate analysis
Training data sets	400	346	400	346	400
Test data sets	100	93	100	93	100
Number of input variables	27	25	25	23	5
Multinomial logistic regression in Python-based TensorFlow					
Training accuracy (%)	99.00	98.84	84.00	79.19	86.75
Test accuracy (%)	93.00	92.50	82.00	81.70	94.00
Precision (%)	95.00	94.00	80.00	79.00	93.00
Recall (%)	93.00	92.00	82.00	82.00	94.00
F1 score (%)	93.00	93.00	80.00	79.00	93.00
Multinomial logistic regression in R					
Test accuracy (%)	97.00	93.50	81.00	80.60	94.00
Precision (%)	97.00	95.00	79.00	78.00	93.00
Recall (%)	97.00	94.00	81.00	81.00	94.00
F1 score (%)	97.00	94.00	79.00	78.00	93.00
Multiclass logistic regression in MAMLS					
Test accuracy (%)	100	97.80	78.00	81.70	93.00
Precision (%)	100	98.00	74.00	81.00	87.00
Recall (%)	100	98.00	78.00	82.00	93.00
F1 score (%)	100	98.00	74.00	77.00	90.00
Multiclass neural network in MAMLS					
Test accuracy (%)	99.00	95.70	77.00	73.10	94.00
Precision (%)	99.00	96.00	78.00	74.00	93.00
Recall (%)	99.00	96.00	77.00	73.00	94.00
F1 score (%)	99.00	96.00	78.00	73.00	93.00

VE variables related to epiphora, *VDCG* variables related to dacryocystography, *MAMLS* Microsoft azure machine learning studio

that these are useful predictive models. We have found that ML is a more useful predictive model using various data including clinical variables, DCG, and LS rather than CNN which does supervised learning with only LS images.

The presence of epiphora and stenosis in DCG was obtained as important variables in both multivariate analysis and variable importance. Based on this, we developed four sets of input data, depending on whether VE or VDCG is included. In addition, we also developed the data set consisting of the variables obtained through multivariate analysis to reduce the confounding effect among the 27 independent variables.

The patients with epiphora were diagnosed almost completely using ML tools in model 1. For all input variables included in model 1, Python-based TensorFlow, R, and MAMLS had at least 93.00% test accuracy and 93.00% F1 score. The best result was achieved through multiclass logistic regression in MAMLS; the test accuracy was 100%, and F1 score was 100% in model 1. In the case of Python-based

Table 3 Performance comparison of Python-based TensorFlow, R, MAMLS, and nuclear medicine physician in model 4

	Multinomial logistic regression in Python-based TensorFlow	Multinomial logistic regression in R	Multiclass logistic regression in MAMLS	Multiclass neural network in MAMLS	Nuclear medicine physician
Test accuracy (%)	81.70	80.60	81.70	73.10	80.60
Precision (%)	79.00	78.00	81.00	74.00	78.00
Recall (%)	82.00	81.00	82.00	73.00	81.00
F1 score (%)	79.00	78.00	77.00	73.00	75.00

MAMLS Microsoft Azure Machine Learning Studio

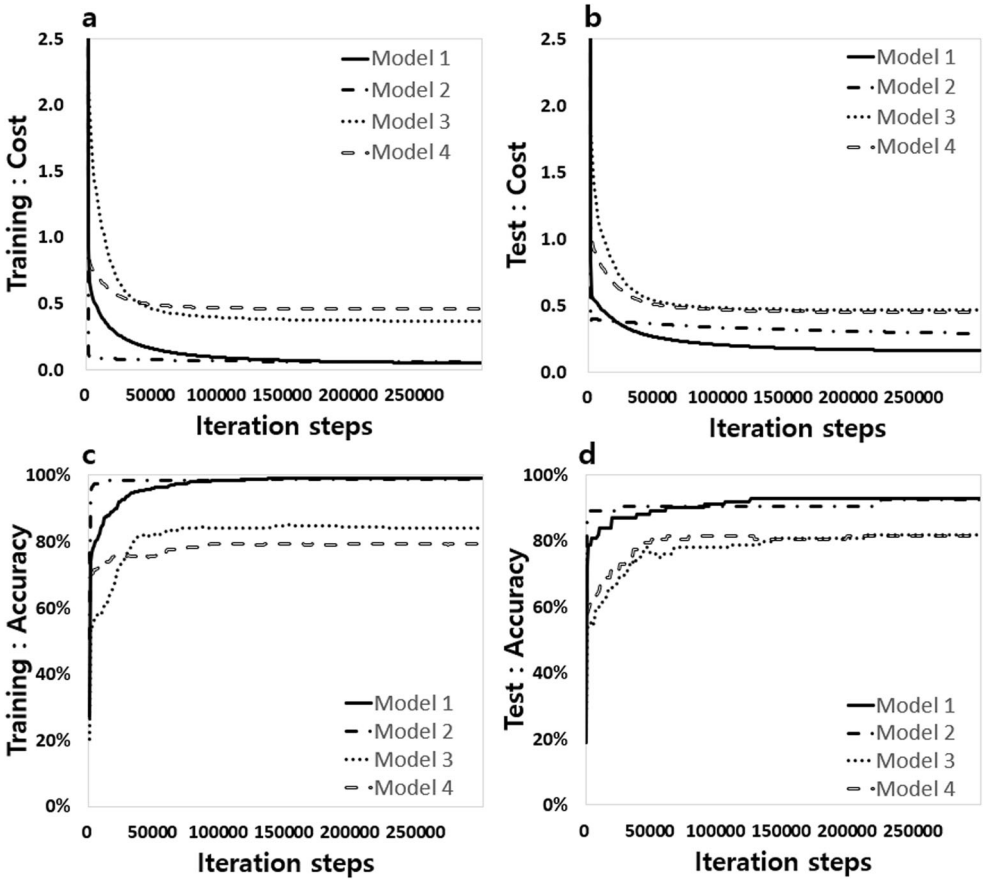

Fig. 5 Costs of training data sets (**a**), costs of test data sets (**b**), accuracies of training data sets (**c**), and accuracies of test data sets (**d**) over iteration steps in Python-based TensorFlow from model 1 to model 4

TensorFlow, the accuracy was good as other ML tools in the training dataset, but in the test dataset, the accuracy was slightly decreased. It may to be due to overfitting. One of the reasons for the better test accuracy of multiclass logistic regression in MAMLS compared to the other cases was probably that regularization was applied. Regularization refers to the method of controlling the model, with the aim of preventing overfitting the model to the training data and hence improving its generalization ability with new data set [12].

We confirmed that the test accuracy of the nuclear medicine physician was similar to the test accuracies of Python-based TensorFlow, R, and multiclass logistic regression in MAMLS in model 4. It has been shown that ML can achieve as much performance as physicians already do [13–15]. Conversely, we found that statistical differences were observed between the multiclass neural network in MAMLS and the remaining ML models. Supervised learning of the multiclass neural network in MAMLS did not work well compared to the other models in this study. This result is thought to be due to overfitting which is a common problem in data science. Overfitting refers to a phenomenon in which a model exactly corresponds within the training data set, but not within the test data set. An overfitted model could not be generalized well from the training data set to additional data sets [16]. If an overfitted model is applied to new test data set from the training data set, it has the potential to produce erroneous predictions [17]. Overfitted model selection is likely to be very severe when the training data set is small, and the number of hyperparameters to be adjusted is relatively large. It is important to avoid overfitting of the model during training iteration; this can be achieved by regularization, early stopping, and hyperparameter tuning [12, 18, 19]. The multiclass neural network in MAMLS used fully connected network; we could have achieved higher test accuracy and reduced overfitting if we used dropout. Dropout is one of the methods of regularization for decreasing overfitting by preventing complex co-adaptations on training data set [20].

We compared model 1 with all 27 independent variables and model 5 with five variables obtained after multivariate analysis. In Python-based TensorFlow, training accuracy in model 5 was lower than model 1, but the test accuracy in model 5 was higher than model 1. This means that learning efficiency of model 5 is better than model 1. However, we confirmed that both test accuracy and F1 score of model 1 were higher than model 5 in R and MAMLS. Also, AUCs in functional obstruction in model 1 were obtained higher than AUCs in functional obstruction in model 5 in ML tools in Supplementary Table 1. Multivariate feature selection has

the advantage of being fast, but is not reliant on any algorithm. However, this method may reduce the classification accuracy due to disregarding communication with the classifier [21].

In this study, we used three ML tools: Python-based TensorFlow, R, and MAMLS. Python-based TensorFlow and R are open-source software libraries for numerical computation and statistical computing, and are widely used for ML and DL. Python-based TensorFlow and R allow users to create their own models through coding and are relatively easy to use on a personal computer. Python-based TensorFlow and R make it easy to change and optimize the various conditions of the model as desired by the developer. However, if we do not know how to code or know the basics of algorithms, using Python-based TensorFlow and R can be difficult. On the other hand, MAMLS, a commercial computing platform made by Microsoft, unlike Python-based TensorFlow and R, is available only in the cloud and easy to use even by people who are not familiar with computer programming. Between TensorFlow and R, R was more superior in convenience. In TensorFlow, hypothesis function, cost function, learning rate, and number of iterations are entered in the learning algorithm, but this process was not required in R package. In addition, the calculation time in R was shorter than in Python-based TensorFlow. A variety of ML algorithms are available through classification models in MAMLS. The existing ML modules are modified and combined with modules in the R language to achieve a generalized flow, which can be acquired from a wide range of applications. Maximized multiclass and binary classification accuracies with minimal manual intervention are achievable in this platform.

Test accuracies of CNN tended to be lower than that of ML. It is probably caused by the input data sets of CNN and ML. While CNN was learned using only LS images, ML models were learned using various data including clinical variables, DCG, and LS, which probably make more accurate predictions. Most of all, LS may be abnormal in up to 40% of asymptomatic eyes with no tracer going down the nose in 25–32% [10]. Therefore, it is not surprising that ML with image and clinical variables had better test accuracies than CNN with only LS images which has insufficient accuracy.

This study has several limitations. The retrospective nature and small number of study patients are problems. The predictive power of the classifiers is largely dependent on the quality and size of the training sample [22]. In this study, the number of patients with epiphora was insufficient for developing a precise predictive model; thus, a larger scaled study should be done in the future. A multicenter study is a considerable option because it allows for a sufficient number of diverse participants in a substantially shorter period of time.

Conclusion

This study is the first study to develop predictive models using LS, ML, and DL. ML-based predictive models classified the patients with epiphora well rather than CNN, and the performance of these models is comparable to the nuclear medicine physician. However, the multiclass neural network model in MAMLS was inferior to other predictive models and the nuclear medicine physician, which may be largely due to overfitting. Although currently available programming languages and computing platforms for ML and DL, Python-based TensorFlow, R, and MAMLS, each have advantages and disadvantages; these languages or platforms have sufficient potential to make a ML and DL model available even for physicians who do not have a high level of expertise in computer programming. These techniques may accelerate the application of clinical diagnoses classification using ML and DL to be faster and wider than we anticipate in the field of nuclear medicine.

Compliance with Ethical Standards

Conflict of Interest Yong-Jin Park, Ji Hoon Bae, Mu Heon Shin, Seung Hyup Hyun, Young Seok Cho, Yearn Seong Choe, Joon Young Choi, Kyung-Han Lee, Byung-Tae Kim, and Seung Hwan Moon declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of our institutional review board and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent The institutional review board of our institute approved this retrospective study, and the requirement to obtain informed consent was waived.

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

References

1. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18:570–84.
2. Choi H. Deep learning in nuclear medicine and molecular imaging: current perspectives and future directions. Nucl Med Mol Imaging. 2018;52:109–18.
3. Fau KK, Wu D, Fau WD, Gong K, Fau GK, Dutta J, et al. Penalized PET reconstruction using deep learning prior and local linear fitting. IEEE Trans Med Imaging. 2018;37.
4. Choi JY. Radiomics and deep learning in clinical imaging: what should we do? Nucl Med Mol Imaging. 2018;52:89–90.
5. Muhlestein WE, Akagi DS, Kallos JA, Morone PJ, Weaver KD, Thompson RC, et al. Using a guided machine learning ensemble

model to predict discharge disposition following meningioma resection. J Neurol Surg B Skull Base. 2018;79:123–30.
6. Papp L, Potsch N, Grahovac M, Schmidbauer V, Woehrer A, Preusser M, et al. Glioma survival prediction with combined analysis of in vivo (11)C-MET PET features, ex vivo features, and patient features by supervised machine learning. J Nucl Med. 2018;59:892–9.
7. Juarez-Orozco LE, Knol RJJ, Sanchez-Catasus CA, Martinez-Manzanera O, van der Zant FM, Knuuti J. Machine learning in the integration of simple variables for identifying patients with myocardial ischemia. J Nucl Cardiol 2018.
8. Betancur J, Otaki Y, Motwani M, Fish MB, Lemley M, Dey D, et al. Prognostic value of combined clinical and myocardial perfusion imaging data using machine learning. J Am Coll Cardiol Img 2017.
9. Vonica OA, Obi E, Sipkova Z, Soare C, Pearson AR. The value of lacrimal scintillography in the assessment of patients with epiphora. Eye (Lond). 2017;31:1020–6.
10. Sagili S, Selva D, Malhotra R, Malhotra R. Lacrimal scintigraphy: "interpretation more art than science". Orbit. 2012;31:77–85.
11. Cuthbertson FM, Webber S. Assessment of functional nasolacrimal duct obstruction—a survey of ophthalmologists in the southwest. Eye (Lond). 2004;18:20–3.
12. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 2014;10:e1004754.
13. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69:2657–64.
14. Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. Comput Biol Med. 2017;82:80–6.
15. Gibbons C, Richards S, Valderas JM, Campbell J. Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. J Med Internet Res. 2017;19:e65.
16. Sun B, Lam D, Yang D, Grantham K, Zhang T, Mutic S, et al. A machine learning approach to the accurate prediction of monitor units for a compact proton machine. Med Phys. 2018;45:2243–51.
17. Pan L, Cheng C, Haberkorn U, Dimitrakopoulou-Strauss A. Machine learning-based kinetic modeling: a robust and reproducible solution for quantitative analysis of dynamic PET data. Phys Med Biol. 2017;62:3566–81.
18. Fontenla-Romero O, Perez-Sanchez B, Guijarro-Berdinas B. LANN-SVD: a non-iterative SVD-based learning algorithm for one-layer neural networks. IEEE Trans Neural Netw Learn Syst 2017:1–6.
19. Yao Y, Rosasco L, Caponnetto A. On early stopping in gradient descent learning. Constr Approx. 2007;26:289–315.
20. Baldi P, Sadowski P. The dropout learning algorithm. Artif Intell. 2014;210:78–122.
21. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinforma. 2015;2015:198363.
22. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 2012;12:8.