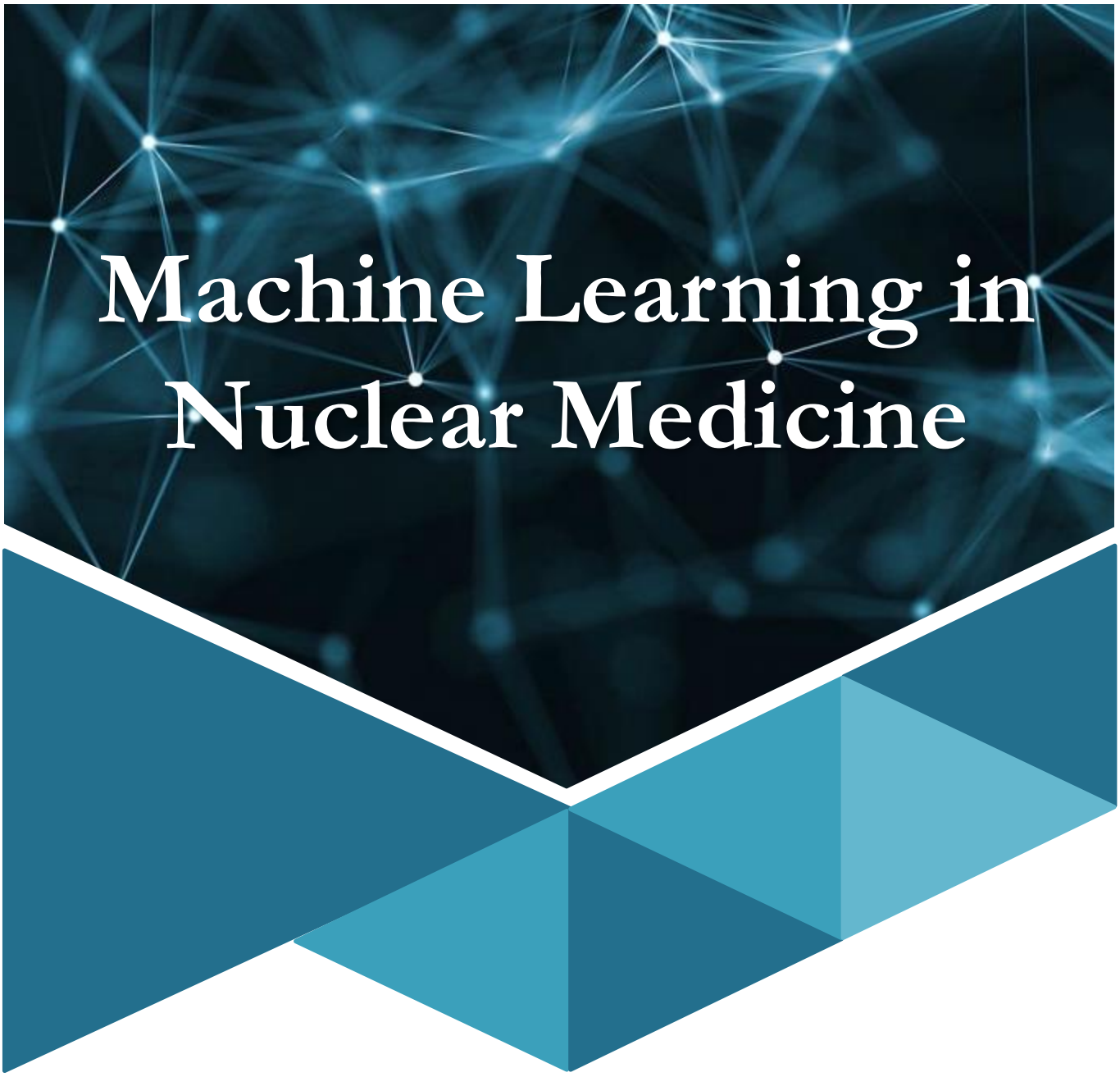


Machine Learning in Nuclear Medicine

Detection of Parkinson's disease, myocardial ischemia and major adverse risk events using support vector machines

UNIVERSITY OF TWENTE.

Maaike Dotinga

Master thesis Technical Medicine

August 2019

MACHINE LEARNING IN NUCLEAR MEDICINE

DETECTION OF PARKINSON'S DISEASE, MYOCARDIAL ISCHEMIA AND MAJOR ADVERSE CARDIAC EVENTS USING SUPPORT VECTOR MACHINES

Author: Maaike Dotinga

Master thesis for the degree of Master of Science in Technical Medicine, Medical Imaging and Interventions. This thesis symbolizes the completion of my clinical specialization internship at the Nuclear Medicine department of Isala Hospital in Zwolle.

Defense:
30 August 2019, 13:00
Carré 2K, University of Twente

Graduation comittee:

dr. Joris D. van Dijk, MSc.	drs. Bregje J.C.C. Hessink-Sweep
dr. Jorn A. van Dalen	prof. dr. ir. Cornelis H. Slump
drs. Brian N. Vendel	prof. dr. ir. Ruud M. Verdaasdonk

SUMMARY

The use of machine learning (ML) in medical imaging is rapidly increasing. Its ability of mapping the relationship between a defined input and output through learning by examples makes ML useful for a wide range of applications in radiology and nuclear medicine. As multiple parameters can be integrated into a single outcome, it particularly has great potential to enhance diagnosis and prognosis to facilitate patient management. In this thesis, we focus on the use of support vector machines (SVMs) to detect Parkinson's disease (PD), myocardial ischemia and major adverse cardiac events (MACE).

In the first study, we aimed to develop and validate a linear SVM classifier to discriminate PD from non-PD based on I-123 FP-CIT SPECT striatal uptake ratios, age and gender. Its generalizability was assessed using previously unseen datasets from two centers using comparable acquisition and image processing methods, thereby comparing prediction performances of the derived model between sites. Expert nuclear medicine physicians scored I-123 FP-CIT SPECT scans of both datasets as either PD or non-PD after which their prediction performance was compared to that of the SVM model to assess its clinical value. We found that comparable prediction performances were obtained between sites with classification accuracies of 95.0% for the dataset from the same center the model was developed and 82.5% for the other. The performances found were similar to that of nuclear medicine physicians who achieved accuracies of 95.0% and 81.3%, respectively. Using the derived SVM, accurate discrimination of PD from non-PD can be achieved that is equivalent to standard visual assessment by expert nuclear medicine physicians. Furthermore, we can assume that the model is generalizable towards centers using comparable acquisition and image processing methods. Hence, implementation of this SVM model as diagnostic aid in clinical practice is encouraged.

In the second study, two Gaussian SVM classifiers were developed to identify patients with myocardial ischemia and patients at risk of MACE. The derived models were subsequently validated using a previously unseen dataset. Input features included various clinical parameters and coronary artery calcium score (CAC) for both models and for the MACE model, left ventricular ejection fraction and myocardial perfusion imaging (MPI) SPECT scan outcome were added. The ischemia model was further evaluated by comparing its prediction performance to that of absent CAC indicative of the absence of ischemia. Validation of the ischemia model led to a sensitivity and specificity of 89.7% and 31.8%, respectively. A comparable prediction performance was found for predicting ischemia using absent CAC, obtaining a sensitivity of 84.1% and specificity of 37.1%. We observed that the MACE model was not generalizable towards previously unseen data as specificity decreased substantially from 16.5% to 3.1%. The higher amount of input features needed to predict ischemia in comparison to standalone CAC scoring and the low specificity of the MACE model impede clinical implementation of these models. Further evaluation of both models is therefore needed to be able to provide a more individualized risk assessment of ischemia and MACE using ML.

DANKWOORD

Een scriptie is niks zonder dankwoord. Bij alles wat ik het afgelopen jaar heb mogen doen en doorstaan stonden zoveel mensen naast mij. Zowel in het ziekenhuis als op de universiteit, dichtbij huis en verder weg - wat heb ik dit gewaardeerd en wat ga ik dit koesteren. Dankjewel allemaal!

Joris, Jorn en Brian, ik had het me niet anders willen wensen. Bedankt voor al het vertrouwen dat jullie me gaven, de goede gesprekken die er zijn geweest en jullie inhoudelijke en persoonlijke begeleiding - hoe subtiel dan ook. Jullie betrokkenheid is van onschatbare waarde. Ik vond het ontzettend leuk om met jullie in deze nieuwe onderzoekslijn te duiken en er wat proberen van te maken. In één (Friese) zin uitgedrukt: ik bin sa wiis mei jimme!

Bregje, heel erg bedankt voor jouw begeleiding deze afgelopen twee jaar, fijn dat je er altijd was! Mede dankzij jouw scherpe vragen, opmerkingen en inzichten sta ik nu hier heel sterk in mijn schoenen en dat is me ontzettend veel waard. Kees, grote dank voor de input vanuit jouw perspectief. Ik vind het knap dat je als technisch begeleider de kliniek erbij blijft proberen te betrekken, waardeer ik echt heel erg. Beste Ruud, dank dat je als buitenlid in mijn afstudeercommissie wilde aanschuiven.

Lieve collega's van de afdeling Nucleaire Geneeskunde en het Klinisch Chemisch Lab, wat zijn jullie betrokken met elkaar en wat ben ik blij dat ik deel van jullie team mocht uitmaken. Bedankt voor jullie hulp tijdens mijn onderzoek en klinische activiteiten, maar vooral ook heel erg bedankt dat jullie me zo thuis kunnen laten voelen!

Lieve familie, vrienden, oud-huisgenootjes, medestudenten, orkestgenootjes of een combi daarvan - een dikke dankjewel naar jullie allemaal die in de afgelopen jaren Technische Geneeskunde iets voor mij heeft betekend. Het is heerlijk om de boel de boel te laten en gewoon lekker bij jullie te zijn. Hopelijk komt daar snel een uitnodiging om bij mij langs te komen op mijn nieuwe stek en kunnen we fijn uitwaaien aan zee ☺

Ik wens jullie allemaal het allerbeste!

Liefs,
Maaike

Zwolle, 17 augustus 2019

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AUC	Area under the curve
BP	Blood pressure
CAC	Coronary artery calcification
CAD	Coronary artery disease
CAG	Coronary angiography
CCTA	Coronary computed tomography angiography
CI	Confidence interval
CT	Computer tomography
CV	Cross-validation
CZT	Cadmium zinc telluride
DAT	Dopamine transporter
DM	Diabetes mellitus
FP-CIT	N-ω-fluoropropyl 2β-carbomethoxy-3β-(4-iodophenyl)nortropane
ICC	Intra-class correlation coefficient
LBBB	Left bundle branch block
MACE	Major adverse cardiac events
ML	Machine learning
MPI	Myocardial perfusion imaging
MFR	Myocardial flow reserve
MR	Magnetic resonance
NLP	Natural language processing
NPV	Negative predictive value
PD	Parkinson’s disease
PET	Positron emission tomography
PPV	Positive predictive value
ROI	Region of interest
SPECT	Single photon emission computer tomography
SN	Substantia nigra
SBR	Specific binding ratio
SVM	Support vector machine
SWEDD	Scans without evidence for dopaminergic deficit

CHAPTER 1

General introduction

Clinical specialists are currently facing a rapid increase of biomedical data, including increasing numbers and complexity of medical images.[1–4] Interpretation of these images to detect, characterize and monitor diseases is merely done by visual assessment that is time-consuming and prone to subjectivity. This can lead to interpretation errors and overlooking of findings that can have a significant clinical impact on patient management and outcome.[5, 6] Advanced image analysis can help to enhance consistency of image interpretation independent of reader experience.[7] In particular, machine learning (ML) is increasingly used in radiology and nuclear medicine to handle the heterogenous and multidimensional nature of medical images and decipher its clinical meaning.[8–10]

ML is a subfield of artificial intelligence that refers to algorithms that learn from data without being explicitly programmed.[11] These algorithms aim at finding and recognizing patterns in (medical) data, thereby computing which features are important in a particular classification or regression problem. By learning from training data, the algorithm optimizes its parameters to improve its performance. The eventually developed model can then be applied to unseen data to generate a prediction of interest.[9, 12] Patterns in data are recognized and captured in order to discriminate data for applications ranging from disease classification and risk stratification to image processing as segmentation and quantification.[13–16] The emerge of deep learning further encourages a more widespread use of ML in medical imaging. [6, 17] Whereas ML involves the design of handcrafted features that requires expert knowledge, DL can automatically discover representations of input data for a specific task as part of its search process.[11, 18]

ML is mainly used in anatomical computed tomography (CT) and magnetic resonance (MR) images, though its application in molecular imaging has high potential to perform automated quantitative analysis for functional imaging. As we aim to represent and visualize biological processes underlying physiologic and pathologic changes in molecular imaging, key pathophysiologic processes can be potentially discriminated by summarizing high-dimensional data into a single outcome parameter.[8] ML-based image interpretation is a promising technique to aid in clinical decision making and to increase speed, accuracy and reproducibility of interpretation, thereby overcoming undesired variation including inter-observer variability.[19–21] Furthermore, integrating imaging data together with clinical factors using ML has the potential to enable more accurate diagnosis and assessment of individual patient outcome.[1, 5, 6, 22]

CLINICAL BACKGROUND

Parkinson's disease

Parkinson's disease (PD) is the second most common neurodegenerative disorder worldwide. Approximately 1% of the population above 60 years is known to have PD, increasing to 5% of the population above 85 years.[23, 24] PD causes considerable disability and reduction of quality of life and imposes a significant economic and social burden that is expected to continue to increase with the ageing of the population.[25–27] The disease is characterized by the pathophysiologic loss of dopaminergic neurons in the substantia nigra (SN) due to widespread intracellular protein (α-synuclein) accumulation, resulting in striatal dopamine deficiency. As dopamine plays a key role in the control of actions and goal-directed behavior, a consequent impairment of motor function will be developed.[28, 29] Typically, patients experience motor manifestations after 50-60% of neurons in the

SN are lost. Cardinal signs include bradykinesia, rest tremor and rigidity.[23, 30, 31] As the disease progresses, motor symptoms worsen, previously unilateral signs become bilateral and additional signs as cognitive impairment, behavioral changes and symptoms related to autonomic nervous system failure become evident.[29, 31]

The cause of PD remains unknown and validated biomarkers to detect and monitor the disease are currently lacking.[29, 32] Hence, diagnosis of PD is a clinical diagnostic decision, mainly based on symptoms and clinical findings as presence of cardinal signs and non-motor features, gradual symptom progression and response to pharmacological dopamine substitution (levodopa) treatment.[29, 33] Diagnostic accuracy is still suboptimal and varies considerably according to disease duration, reaching an accuracy of approximately 84% in a specialist setting.[32, 34] Particularly in early stages of PD, accurate diagnosis is key to accommodate proper patient management.[35, 36] Visualizing neuroanatomical features with MR imaging and functional signatures by means of dopamine transporter (DAT) imaging with single photon emission computer tomography (SPECT) can provide improvements in diagnostic accuracy. [31, 37] MR imaging helps to rule out secondary causes of parkinsonism as multisystem atrophy, but the morphological alterations are usually detectable at advanced stages and this technique is limited to use for identifying the dopaminergic deficits that are characteristic in PD.[38–40] As an early response to decreased synaptic dopamine concentration, DATs are known to be downregulated in PD which can be evaluated by visualizing the uptake of dopaminergic tracers. In Europe, I-123 N-ω-fluoropropyl 2β-carbomethoxy-3β-(4-iodophenyl) nortropane (FP-CIT) is the most widely used tracer for DAT imaging that is approved for clinical use by the Food and Drug Administration.[36, 41] Using this technique, the loss of dopaminergic neurons in the SN can already be detected in early stages of PD by DAT imaging as the uptake of dopaminergic tracers reflects the integrity of these neurons.[36, 42] Unfortunately, parkinsonian syndromes beyond PD and their overlap in clinical, pathological and imaging features are increasingly recognized.[32, 43] Hence, correct identification of PD still poses a considerable challenge for clinicians in which ML has the potential to help.

Coronary artery disease

Coronary artery disease (CAD) is the most common cause of death in both men and women, accounting for approximately 20% of all deaths in Europe.[44] CAD comprises a reduction of blood flow to the myocardium, often due to atherosclerosis, that involves narrowing of the arteries caused by the build-up of plaque. Consequently, there is an imbalance between myocardial oxygen supply and demand which often manifests as angina pectoris.[45–47] If CAD is left untreated, progression of the disease can lead to myocardial infarction, heart failure and eventually sudden cardiac death.[48, 49] Early detection and accurate diagnosis of CAD is therefore essential for patient management, thereby avoiding unnecessary invasive and non-invasive tests and treatments.[47, 50] Anatomical and functional imaging of the myocardium and coronary arteries is pivotal in the evaluation of patients suspected for CAD as these techniques can provide information about the location, extent and functional severity of the lesion.

The definitive diagnosis of significant CAD is commonly made during coronary angiography (CAG), but ideally reserved for patients likely to require revascularization due to its invasive nature.[51, 52] Non-invasive techniques including coronary computer tomography angiography (CCTA) and myocardial perfusion imaging (MPI) are widely used prior to invasive CAG for evaluating CAD. Cardiac CT allows the detection of stenosis and assessment of the total atherosclerotic burden, while MPI with SPECT or positron emission tomography (PET) can evaluate the presence of and extent of perfusion defects in rest and stress conditions. Both have significant prognostic value and aid in the

identification of patients that might benefit from invasive interventions besides medical treatment, including percutaneous coronary intervention and coronary artery bypass grafting.[53–55]

Particularly, the presence and extent of myocardial ischemia is a key factor in the management of CAD.[56] This can be detected by MPI, thereby evaluating the uptake of radioactive tracers into the myocardium at rest and stress. In ischemic areas, the decreased blood flow results in a reduction in myocardial uptake that can be observed on SPECT and PET scans.[51] As coronary atherosclerosis is often accompanied with calcium deposits, assessment of the coronary calcified plaque burden can be used to stratify the risk of CAD.[57, 58] This is commonly done by measuring the coronary artery calcium (CAC) score with non-contrast CT. Using the Agatston scoring methodology, the volume of calcium in all coronary arteries and density of each calcified plaque is taken into account to provide a single score that reflects the atherosclerotic burden.[51, 59] The combination of imaging results with clinical factors and demographic features has the potential to refine current identification of patients at risk.[60, 61] Using advanced ML algorithms, a more individual risk stratification approach can be developed that facilitates decision making and help in the overall management of patients with known or suspected CAD.

TECHNICAL BACKGROUND

Machine learning

ML refers to algorithms that are able to automatically model the relationship between a given input and output. These algorithms are programmed to learn from observations to make statistical inferences that are used to generate predictions for novel input values.[62, 63] Three main types of machine learning are distinguished; supervised learning, unsupervised learning and reinforcement learning. In supervised learning, a training set of examples with target labels or responses is provided that is used to train the model. The aim of these models is to assign an input vector to a discrete category or to predict a continuous variable, having either a classification or regression task, respectively.[64, 65] In unsupervised learning, the algorithm receives an unlabeled set of data inputs and tries to find similarities in the input data for clustering and dimensionality reduction purposes.[63, 66] Reinforcement learning comprises learning from experience coupled with the principle of reward and punishment. The algorithm is provided feedback on its decisions until a proper solution is found for a specific task.[63, 67]

In this thesis, we focus on the use of supervised learning in binary classification problems. In this case, the training dataset comprises input vectors $\{x_1, \dots, x_n\}$ with corresponding output values $\{y_1, \dots, y_n\}$ where $y_i \in \{0,1\}$. In the training phase, the algorithm learns from these examples to tune parameters or weights of the model such that the rate of misclassification decreases. Eventually, a function $y(x)$ is determined after which the ability of categorizing previously unseen examples, known as generalization, is assessed.[64, 65] For this, a labelled test set is used to compare predicted output with the label in order to determine how well the algorithm has learned.[62] The generalizability of a model is dependent on how a model is fitted to the training dataset. An inaccurate model is not matching the data well due to either the incapability of capturing the underlying function (bias) or to sensitivity to noise in the data (variance). This can lead to either underfitting, comprising a high bias that leads to a large error in both training and test set or to overfitting, where a large variance is seen resulting in a poor fit for novel data in the test set as illustrated in Figure 1.

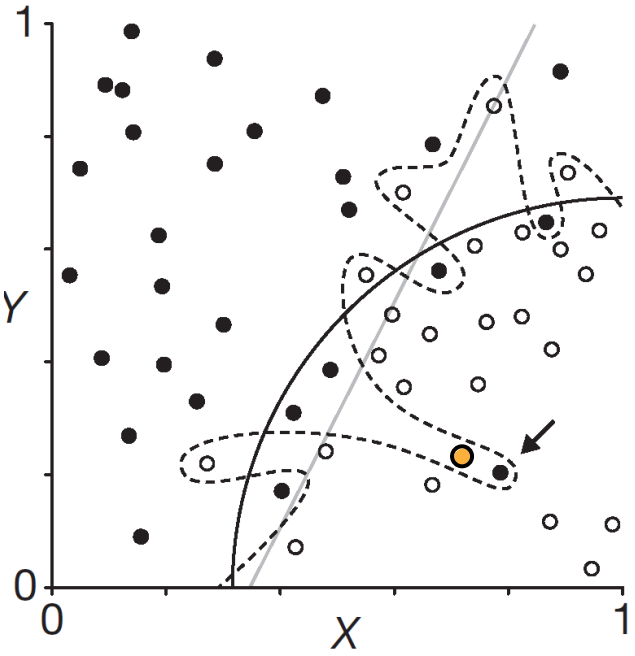


Figure 1: Binary classification problem in which open and solid circles have to be discriminated. A reasonable separation line (black curve) is illustrated as well as decision boundaries in case of underfitting (gray diagonal line) and overfitting (dashed curve). The overfit is caused by an outlier (arrow) which would result in the misclassification of a novel point (orange circle).[68]

Model complexity, comprising model type, number of inputs and number of parameters, is known to influence bias and variance. Therefore, proper model selection is needed, thereby choosing the right parameters for the model so that it generalizes as well as possible.[62] To evaluate multiple models and parameter values, without the need for additional test sets, a cross-validation (CV) can be performed. Most used approaches include the *k*-fold CV and leave-one-out CV. In case of *k*-fold CV, the data is split into *k* groups of approximately the same size. Then, the algorithm is run *k* times using (*k*-1) groups for training and the remaining group as validation set. One example in which $k = 5$ is illustrated in Figure 2. Leave-one-out CV is a special case of k-fold CV where the number of groups equals the number of examples. At each fold, exactly one example is left out to be tested by the model trained with the other examples. In general, higher values of *k* are more prone to overfitting.[69–71]

Multiple algorithms can be used to solve binary classification problems, including algorithms comprising Bayesian models, decision trees, logistic models, support vector machine (SVM) models and artificial neural networks.[62, 64] Usually, SVMs and neural networks tend to perform better when dealing with multiple features as these algorithms have the ability to handle high-dimensional data.[72, 73] SVM is thereby considered to be easier to use and is known for its good generalization performance.[73, 74] Therefore, we focus on the use of SVM to solve classification problems that are encountered in nuclear medicine. Further details of other ML algorithms will not be explained as these lie beyond the scope of this thesis.

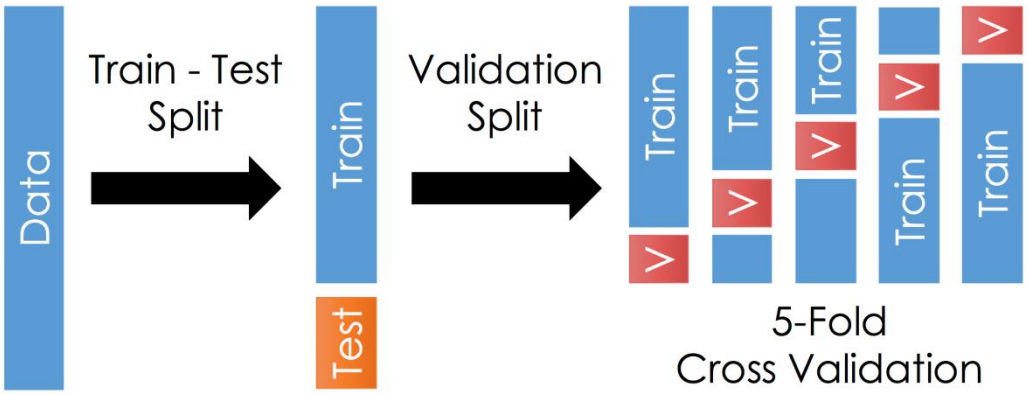


Figure 2: Principle of model evaluation using 5-fold cross-validation and a hold-out test set. The data is split into a training and test group after which the training dataset is divided into five groups. For each fold, one group is used to test the model (red) trained with the remaining groups (blue). The test set (orange) is then used to evaluate generalization of the derived model.

Support vector machines

Support vector machines (SVMs) were developed by Vapnik et al. and are widely used in solving classification and regression problems. In the binary classification problem, the goal of SVMs is to find a decision boundary, or hyperplane, between two classes of data. We can define many possible solutions for this, but SVM searches for the most optimal hyperplane that is able to separate training data with minimal errors and that maximizes the margin between the vectors closest to the hyperplane as illustrated in Figure 3. Only a small amount of the training data is used to construct this hyperplane, the so-called support vectors, which are the points that are the most prone to misclassification. For novel data points, the SVM model can state whether they lie on one side of the hyperplane or the other.[64, 75, 76]

We can distinguish both hard margin and soft margin SVMs. For the hard margin, all points are enforced out of the margin and a model that allows zero errors is fitted to the data. This can lead to overfitting when data contains outliers and is not completely separable. For classes that cannot be perfectly separated, some misclassification is required to be able to construct a hyperplane that is generalizable towards previously unseen data. This can be achieved by using a soft margin, thereby introducing a regularization parameter that controls the trade-off between maximizing the margin and minimizing the error.[64, 77]

SVM is a linear classifier, though nonlinear models can be obtained by transforming the input data using kernels. The so-called kernel trick gives the ability to generate non-linear decision boundaries, thereby making SVM applicable for both linearly and non-linearly separable data.[78] A kernel results in the projection of the data into a higher dimensional feature space after which the data is separated by a linear function. This linear function in the feature space corresponds to a non-linear decision boundary in the original input space. The most commonly used kernels are the linear, polynomial, sigmoid and Gaussian kernel and choosing the most appropriate kernel highly depends on the problem at hand. Apart from the regularization parameter, the decision boundary and thus the performance of the model is dependent on the kernel and its adjustable parameters (e.g. degree of polynomial kernel and width of a Gaussian kernel).[62, 72, 77] The effect of increasing values of these hyperparameters on the decision boundary is adequately explained and illustrated by Ben-Hur et al. [72] Appropriate values of the parameters are found by experimenting with different values, commonly performed in a

systematic manner by grid-search. The grid points are generally chosen on a logarithmic scale and for each combination of parameters, classifier accuracy is determined.[72]

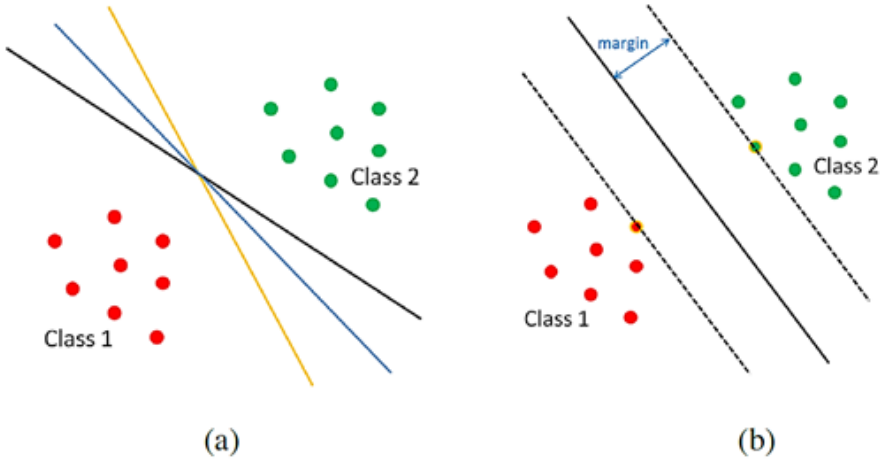


Figure 3: (a) Possible decision boundaries for the binary classification problem. (b) Hyperplane with maximum margins between classes.[79]

AIM AND OUTLINE OF THE THESIS

In the studies presented in this thesis, we focus on the use of ML to (1) refine the interpretation of I-123 FP-CIT SPECT scans and (2) to improve patient selection for MPI SPECT and risk assessment of major adverse cardiac events (MACE).

In Chapter 2, we developed and evaluated the prediction performance of a SVM classifier to detect Parkinson's disease using semi-quantitative results of I-123 FP-CIT SPECT scans combined with demographic features. A two-center study was conducted to assess generalizability of the developed classifier towards centers using the same acquisition and image processing methods. The clinical value of the model was further assessed by comparing its performance to that of expert nuclear medicine physicians. In Chapter 3, we first focus on the development and validation of a SVM classifier that discriminates patients with an abnormal SPECT MPI scan from patients with a normal scan. The model integrates readily available demographic and clinical parameters which were previously collected for a large cohort of patients. Its prediction performance was compared to that of absent coronary artery calcification as determined by coronary CT which is currently used for SPECT MPI patient selection. A subset of these patients who were followed up after SPECT MPI and for whom patient outcome was recorded were used to build and validate a SVM model to predict whether MACE is likely to occur or not. Future perspectives, clinical implications and recommendations for further research comprising ML-based applications are discussed in Chapter 4.

CHAPTER 2

Development, validation and clinical value of machine learning-based interpretation of I-123 FP-CIT scans to detect Parkinson's disease: a two-center study

ABSTRACT

Aim: Our aim was to develop and validate a machine learning (ML)-based approach for interpretation of I-123 FP-CIT SPECT scans to discriminate Parkinson's disease (PD) from non-PD and determine its generalizability and clinical value in two centers.

Methods: We retrospectively included 210 consecutive patients that underwent I-123 FP-CIT SPECT imaging (Discovery D670; GE Healthcare) and had a clinically confirmed diagnosis. Linear support vector machine (SVM) was used to build a classification model to discriminate PD from non-PD based on I-123-FP-CIT striatal uptake ratios, age and gender of 90 patients. The model was validated on previously unseen data from the same site where the model was developed (n=40) and from a different site that utilizes comparable acquisition and image processing methods (n=80). Prediction performance was assessed and compared to that of expert nuclear medicine physicians who scored the I-123 FP-CIT scans as either PD or non-PD.

Results: Testing the derived SVM model on the previously unseen dataset from the same site resulted in an accuracy of 95.0%, sensitivity of 96.0% and specificity of 93.3%. This was identical to the classification accuracy of nuclear medicine physicians ($p > 0.99$) who interpreted both I-123 FP-CIT images and striatal uptake ratios. The model was generalizable towards the other center as a similar prediction performance was found ($p>0.10$) with an accuracy of 82.5%, sensitivity of 88.5% and specificity of 71.4%. This was comparable to that of nuclear medicine physicians who achieved an accuracy, sensitivity and specificity of 81.3%, 84.6 and 88.5% ($p>0.5$), respectively.

Conclusion: ML-based interpretation of I-123-FP-CIT scans results in accurate discrimination of PD from non-PD similar to standard visual assessment in both sites, indicating that the derived SVM model is generalizable towards centers using comparable acquisition and image processing methods. Hence, implementation of this SVM model as diagnostic aid in clinical practice is encouraged.

INTRODUCTION

Single positron emission computer tomography (SPECT) with I-123 N-ω-fluoropropyl 2β-carbomethoxy-3β-(4-iodophenyl)nortropane (FP-CIT) allows for visualization of striatal dopamine deficiency due to the loss of dopaminergic neurons which is characteristic of PD.[39, 80] This imaging technique aids in the diagnostic process as it enhances diagnostic confidence, especially in patients with clinically uncertain parkinsonian syndromes.[81, 82] Re-evaluation of diagnosis is evident in up to 35% of these patients and changes in management and treatment are induced in approximately 70%.[81, 83] Current guidelines recommend the combination of visual assessment and semi-quantitative analysis for adequate interpretation of I-123 FP-CIT scans.[42, 84] Semi-quantitative analysis comprises the assessment of radiopharmaceutical-specific uptake in regions of interest (striatum, caudate nucleus and putamen) and non-specific uptake in reference areas as the occipital lobe.[85] The addition of semi-quantification to visual assessment offers an increase in reader confidence and provides superior diagnostic accuracy when compared to standalone visual assessment.[86, 87] However, no universal cutoff values are available to determine whether semi-

quantitative results are normal or abnormal as acquisition, reconstruction and quantification methods are known to influence these striatal uptake ratios.[88, 89]

Currently, a trend is seen towards machine learning-based approaches for automated classification of I-123 FP-CIT scans to improve interpretation.[90, 91] These approaches have the potential to be exploited as diagnostic aid to nuclear medicine physicians, thereby improving interpretation of these scans and overcoming the limitations of semi-quantitative analysis.[91, 92] Yet usage of developed models across centers is limited as focus is given to generalization beyond a training dataset and not to generalization to different sites.[93] The aim of this study is therefore to develop and validate a machine-learning based approach for interpretation of I-123 FP-CIT scans to detect PD for a single set of acquisition and image processing parameters and to determine its generalizability and clinical value in two centers. Furthermore, we wanted to determine whether the classifier can be used among different technologists by calculating the reproducibility of SVM output among assessed input values.

METHODS

Study population

We retrospectively included a consecutive cohort of 210 patients with clinically confirmed diagnosis that underwent I-123 FP-CIT SPECT imaging between 2014 and 2018 in two medical centers in the Netherlands (Isala hospital, Zwolle and Treant Zorggroep, Scheper hospital, Emmen). Patients' diagnoses were assessed by the attending neurologist according to standard diagnostic criteria.[37] Clinical and demographic data including diagnosis, gender and age at the time of SPECT acquisition were retrieved from patients' medical records and patients were labelled as either having PD or a diagnosis other than PD (non-PD). By means of stratified random sampling, the Isala data was split into a group of 90 patients for model development (further referred by group A) and a group of 40 patients for validation (further referred by group B). Data of Scheper hospital (n=80, further referred by group C) was used to assess generalizability of the model and both group B and C were used in the clinical evaluation.

Image acquisition and reconstruction

Patients were instructed to discontinue medication interfering with I-123 FP-CIT binding to dopamine transporters prior to scanning. SPECT studies were carried out according to standard clinical procedure using a dual-head gamma camera (Group A+B: Discovery D670, GE Healthcare; Group C: Infinia Hawkeye, GE Healthcare) equipped with a low energy, high resolution collimator. Three to six hours before SPECT acquisition, patients were intravenously administered 185 MBq of I-123 FP-CIT. For SPECT data acquisition, a 10% energy window centered on the photopeak of I-123 at 159 keV was used. A total of 64 projections over a circular 360° orbit (rotational radius of approximately 13 cm) were acquired on a 128x128 matrix (1.23-1.28 acquisition zoom, 3.45-3.59 mm pixel size) with an overall scanning time of 32 minutes (30 s per projection).

Image reconstruction was performed by filtered back projection using a Butterworth pre-filter (cut-off 0.65 cycles/cm, order 10) and uniform Chang attenuation correction (coefficient 0.11 cm^{-1}). Attenuation correction was based on a variable ellipsoid map that followed the contour of the head, manually defined using thresholding. Images were reformatted into slices in axial, coronal and transversal planes (3.45-3.59 mm slice thickness) and axial slices were reoriented along the acanthomeatal line.

Semi-quantitative analysis

Semi-quantitative analysis was performed using a functional imaging workstation (Xeleris version 4.0; GE Healthcare) to assess specific I-123 FP-CIT binding in the striatum and striatal subregions including both right and left caudate nucleus and putamen. Non-specific binding was assessed using the occipital cortex as reference region. Five pre-defined fixed regions of interest (ROIs) were manually positioned over caudate nucleus, putamen and occipital cortex on three consecutive slices that were selected best representative of the activity and shape of the striatum. An example of a non-PD and PD patient with correct positioning of all predefined regions is shown in Figure 1. Mean counts within the ROIs were determined after which specific binding ratios (SBRs) were obtained according to the formula:

$$SBR = \frac{striatal\ ROI - occipital\ cortex}{occipital\ cortex} \qquad (1)$$

SBRs for both left and right striatum, caudate nucleus and putamen were assessed, resulting in six different ratios: Str_R , Str_L , $Caud_R$, $Caud_L$, Put_R and Put_L. Furthermore, a putamen/caudate index was calculated by dividing the mean counts within the putamen by the mean counts within the caudate nucleus, thereby obtaining ratios $Put_R/Caud_R$ and $Put_L/Caud_L$.

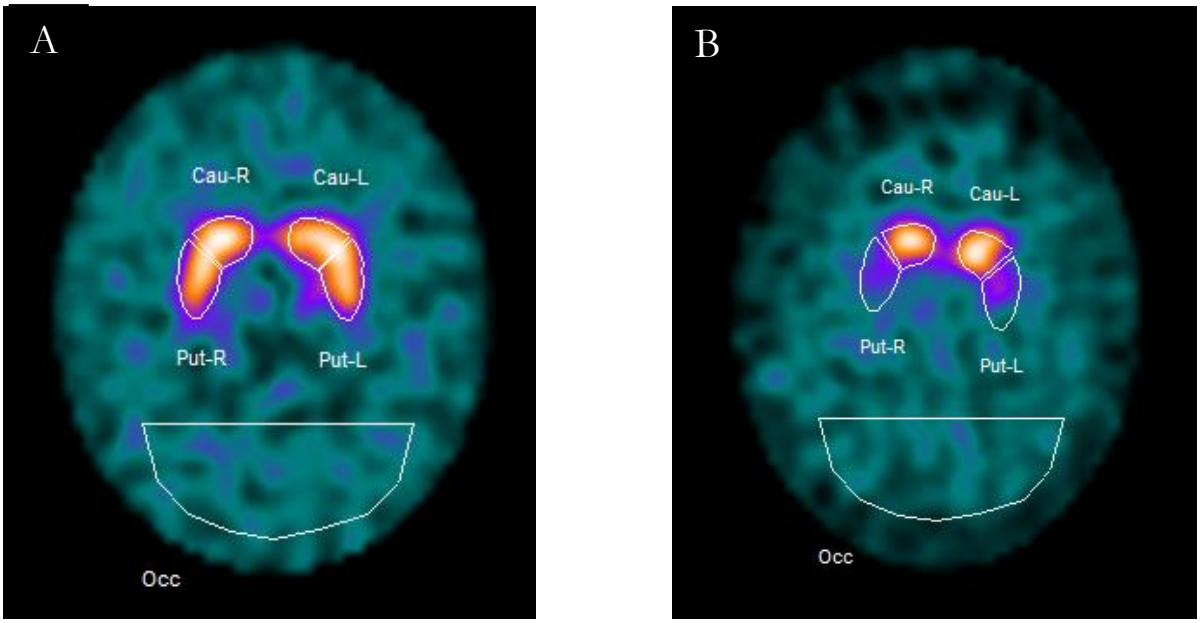


Figure 1: Representative slice of a I-123 FP CIT scan of a non-PD (A) and PD (B) patient with predefined regions for semi-quantitative analysis, including regions for the left and right caudate nuclei (Cau-L and Cau-R) and putamen (Put-R and Put-L) and the occipital cortex (Occ).

SVM classifier development and validation

SVM was used as a classification method to discriminate PD from non-PD based on input features that included I-123 FP-CIT striatal uptake ratios, age and gender. Prior to training and testing procedures, ratios and age were normalized such that the mean value was 0 and standard deviation was 1. All procedures were performed using Matlab software (MATLAB and Statistics and Machine learning Toolbox Release 2018b, the Mathworks Inc.).

The model was built upon ratios assessed by one technologists. Group A was used to build a linear SVM and to perform hyperparameter optimization. A grid search was conducted, thereby

evaluating regularization parameter values of 2^{-6}, 2^{-5}, 2^{-4}, …, 2^{8}. For each value, a stratified, 10-times repeated 10-fold cross-validation was performed after which the mean F1-score was determined. The F1-score is defined as:

$$F_1 = 2 \cdot \frac{positive\ predictive\ value \cdot sensitivity}{positive\ predictive\ value + sensitivity} \quad (2)$$

The value providing the highest mean F1-score was used to derive the final model. This model was further validated by testing group B and C, assessing both predicted class and the probability that test data belongs to PD. For the latter, an inbuilt function for converting SVM scores to probabilities based on logistic regression was used.[94] Given the probabilities, the validation dataset was divided into four categories (<20%, 20-50%, 50-80%, >80% chance of PD). Furthermore, prediction performance was determined by assessing accuracy, F1-score, sensitivity and specificity.

Clinical value

For group B and C, performance of the derived model was compared to that of expert nuclear medicine physicians. Whether and to what extent I-123 FP-CIT images and ratios were typical or characteristic for PD was assessed by two nuclear medicine physicians according to a 4-point scale. This scale consisted of the following categories: unlikely, not probable, probable and certain PD and comprised an expected chance of PD of <20%, 20-50%, 50-80% and >80%, respectively. Images were scored visually, taking into account the magnitude and homogeneity of I-123 FP-CIT distribution, striatal shape and symmetry, definition of striatal borders and the amount of background activity. Images were firstly scored without ratios after which ratios were presented to the nuclear medicine physician and a second score was obtained. In case of disagreement, overread from a third nuclear medicine physician was performed. Images were presented in random order and all readers were blinded to patient characteristics and clinical information except age and gender.

Reproducibility of SVM output

Image reconstruction and semi-quantitative analysis were carried out by one technologist (further referred as observer 1) for all patients. For group B, the reconstruction method as well as assessment of I-123 FP-CIT uptake ratios were repeated by observer 1 and conducted by a different technologist (further referred as observer 2). The SVM classifier was tested using the semi-quantitative analysis results assessed by both observer 1 and 2 after which predicted classes, probabilities and corresponding categories were collected.

Statistical analysis

Statistical analysis was performed using R Studio software (RStudio: Integrated Development for R. Version 1.1.442; RStudio, Inc.). To assess differences between group B and C in age, ratios and gender, the Mann-Whitney U-test or χ^2-test were performed. Accuracies, sensitivities and specificities of the classifier for group A, B and C were compared using the χ^2-test. To determine the reproducibility of the SVM output, the degree of absolute agreement was evaluated by determining the intraclass correlation coefficient (ICC) with corresponding 95%-confidence interval (CI). Prediction performance and scores between observers were compared by a McNemar's test and χ^2-test, respectively. McNemar's test was was also used to compare prediction performance of the SVM model, standalone visual assessment and the combination of visual assessment and ratio interpretation. Differences in frequencies of scores comprising the chance of PD, either determined by the model or scored by nuclear medicine physicians, were assessed using a χ^2-test. A significance level at 0.05 was used and Bonferroni correction was applied when necessary.

Table 1: Patient characteristics including age, male, SBRs and putamen/caudate index per center for PD patients. The p-values are given for either the χ^2-test or Mann-Whitney U-test.

	Group A (n=58)	Group B (n=25)	Group C (n=52)	p-value A vs. B	p-value A vs. C
age	68 (61-74)	67 (60-74)	73 (60-78)	0.7	0.2
male	56.9%	68.0%	69.2%	0.5	0.3
str. L	2.54 (2.28-2.90)	2.73 (2.41-3.08)	2.67 (2.36-2.95)	0.12	0.2
str. R	2.52 (2.35-2.71)	2.61 (2.49-2.87)	2.61 (2.34-3.11)	0.14	0.16
caud. L	3.14 (2.74-3.48)	3.16 (2.95-3.67)	3.32 (2.86-3.68)	0.2	0.16
caud. R	3.07 (2.81-3.29)	3.16 (2.90-3.52)	3.10 (2.81-3.80)	0.2	0.3
put. L	2.04 (1.85-2.37)	2.21 (1.93-2.59)	2.04 (1.82-2.45)	0.09	0.9
put. R	2.05 (1.90-2.25)	2.15 (1.98-2.30)	2.12 (1.82-2.42)	0.3	0.5
put/caud. L	0.67 (0.62-0.74)	0.66 (0.60-0.77)	0.66 (0.57-0.73)	0.8	0.3
put/caud. R	0.69 (0.65-0.76)	0.67 (0.63-0.73)	0.68 (0.62-0.75)	0.3	0.5

Data are presented as median (interquartile range) or percentage; str, striatum; caud, caudate nucleus; put, putamen; L, left; R, right

Table 2: Patient characteristics including age, male, SBRs and putamen/caudate index per center for non-PD patients. The p-values are given for either the χ^2-test or Mann-Whitney U-test.

	Group A (n=32)	Group B (n=15)	Group C (n=28)	p-value A vs. B	p-value A vs. C
age	69 (60-78)	71 (66-77)	74 (69-78)	0.7	0.16
male	37.5%	40%	57.1%	>0.99	0.2
str. L	3.87 (3.44-4.36)	3.76 (3.46-4.07)	3.64 (3.23-4.00)	0.5	0.08
str. R	3.84 (3.31-4.24)	3.78 (3.44-4.01)	3.60 (3.20-3.83)	0.6	0.13
caud. L	4.31 (3.80-4.95)	3.94 (3.57-4.43)	4.08 (3.73-4.47)	0.2	0.3
caud. R	4.07 (3.49-4.64)	4.08 (3.50-4.27)	4.04 (3.63-4.32)	0.6	0.3
put. L	3.60 (3.13-4.13)	3.44 (3.13-4.01)	3.27 (2.58-4.32)	0.8	0.014
put. R	3.55 (3.07-3.56)	3.51 (3.25-3.75)	3.16 (2.96-3.45)	0.8	0.030
put/caud. L	0.83 (0.78-0.89)	0.88 (0.83-0.94)	0.76 (0.70-0.82)	0.10	0.002
put/caud. R	0.87 (0.80-0.90)	0.88 (0.79-0.92)	0.80 (0.73-0.85)	0.7	0.014

Data are presented as median (interquartile range) or percentage; str, striatum; caud, caudate nucleus; put, putamen; L, left; R, right

RESULTS

Patient characteristics

Patient characteristics and different ratios for all PD and non-PD patients are summarized in Table 1 and Table 2, respectively. For PD patients, all variables were comparable between the training dataset and validation datasets from both Isala ($p > 0.09$) and Scheper hospital ($p > 0.16$). Likewise, characteristics and ratios of non-PD patients in group B were similar to group A ($p > 0.2$). No significant differences were found between non-PD patients in group A and C, except for the $Put_L/Caud_L$ index which was significantly lower in group C after Bonferroni correction was applied ($p = 0.002$).

SVM classifier development and validation

A regularization parameter value of 2^{-4} was selected to derive the final model, providing a F1-score of 0.956 ± 0.002 as determined by the 10-times repeated, stratified 10-fold cross-validation. A corresponding accuracy of $94.3\% \pm 0.2\%$ was found for this value. Validation of the derived model lead to accuracies of 95.0% and 82.5% for group B and C, respectively. Prediction performance variables for all groups are shown in Table 3. Prediction performance of the model for group A was comparable to group B ($p > 0.9$) and C ($p > 0.10$). Comparing prediction performance between group B and C, a similar accuracy ($p = 0.3$), sensitivity ($p = 0.5$) and specificity ($p = 0.2$) were found.

Table 3: Prediction performance of the derived SVM model for all groups.

	Group A*	Group B	Group C
Accuracy (%)	94.3 ± 0.2	95.0	82.5
F1-score	0.956 ± 0.002	0.960	0.849
Sensitivity (%)	96.4 ± 0.003	96.0	88.5
Specificity (%)	90.6 ± 0.000	93.3	71.4

*mean values ± SD determined by a 10-times repeated, stratified 10-fold cross-validation

Clinical value

Using the combination of visual assessment and ratio interpretation, nuclear medicine physicians were able to discriminate PD from non-PD with an accuracy of 95% and 81.3% for group B and C, respectively. For both groups, the presentation of ratios besides I-123 FP-CIT images to the physician did not provide an increase in accuracy ($p > 0.5$), sensitivity ($p > 0.99$) or specificity ($p > 0.99$). Additionally, scored chances of PD as assessed by the physician using only visual assessment was comparable to the chances of PD scored using the combination of visual assessment and ratio interpretation ($p > 0.2$). Prediction performances and scored chances of PD of both physicians and the derived SVM model for the two centers are illustrated in Figure 2 and 3, respectively.

Comparing the prediction performance of the SVM model with that of nuclear medicine physicians, equivalent accuracies ($p > 0.4$), sensitivities ($p > 0.3$) and specificities ($p > 0.99$) were found in both groups. In group B, the SVM showed an increase in the confidence of diagnosis, i.e. higher number of scores 1 (<20% chance of PD) and 4 (>80% chance of PD) were found when compared to visual assessment ($p = 0.035$). However, this was not observed in group C ($p = 0.8$). Furthermore, no difference in the frequencies of the scored chance of PD was found

between the SVM model and physicians using the combination of visual assessment and ratio interpretation for both sites ($p > 0.8$).

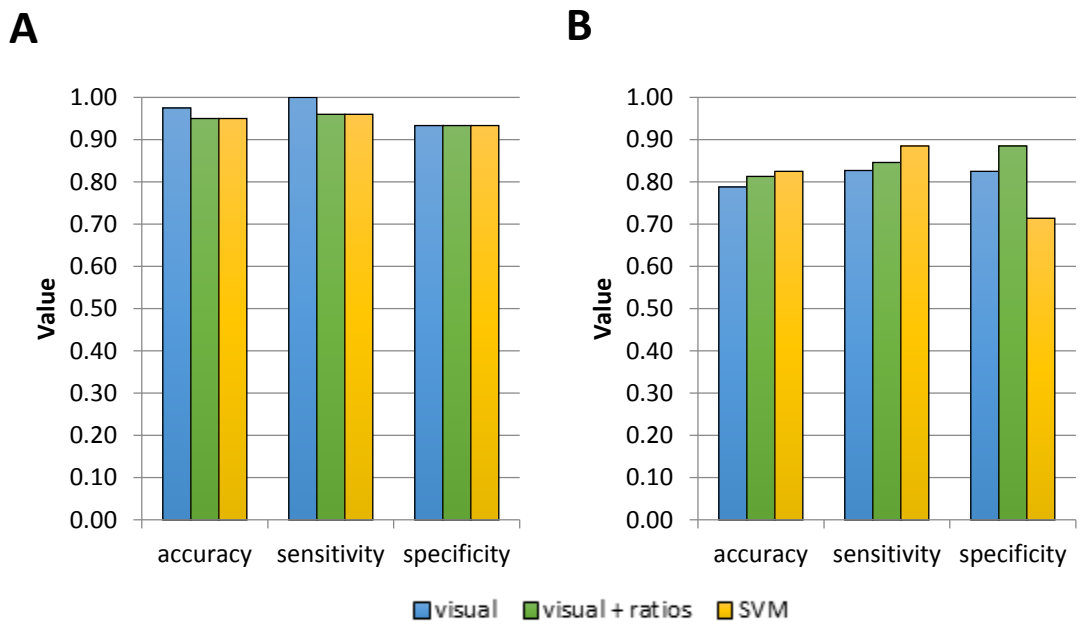


Figure 2: Prediction performance of visual and semiquantitative interpretation by nuclear medicine physicians and interpretation by the SVM model for the validation set of (A) Isala hospital and (B) Scheper hospital. No significant differences were observed between prediction performance variables of the SVM model and nuclear medicine physicians ($p > 0.25$).

Reproducibility of SVM output

The intra-reproducibility of the SVM output was excellent as predicted classes between repeated measures for group A were identical ($p > 0.99$) with an ICC of 0.98 (95%-CI: 0.96-0.99) for assessed probabilities. Additionally, corresponding scores comprising the chance of PD were comparable ($p = 0.6$). Using the ratios assessed by observer 2 as input to the SVM classifier, a comparable accuracy of 90% ($p = 0.5$), sensitivity of 92% ($p > 0.99$) and specificity of 86.7% ($p > 0.99$) was found in comparison to the performance determined using ratios assessed by observer 1 (accuracy, sensitivity and specificity of 95.0%, 96.0% and 93.3%, respectively). A total of two patients (5.0%) that were correctly classified by the SVM model using ratios assessed by observer 1 were misclassified by the model using ratios assessed by observer 2. The inter-reproducibility of the SVM output was excellent, obtaining an ICC of 0.98 (95%-CI: 0.93-0.98) for assessed probabilities and similar corresponding scores comprising the chance of PD ($p = 0.97$).

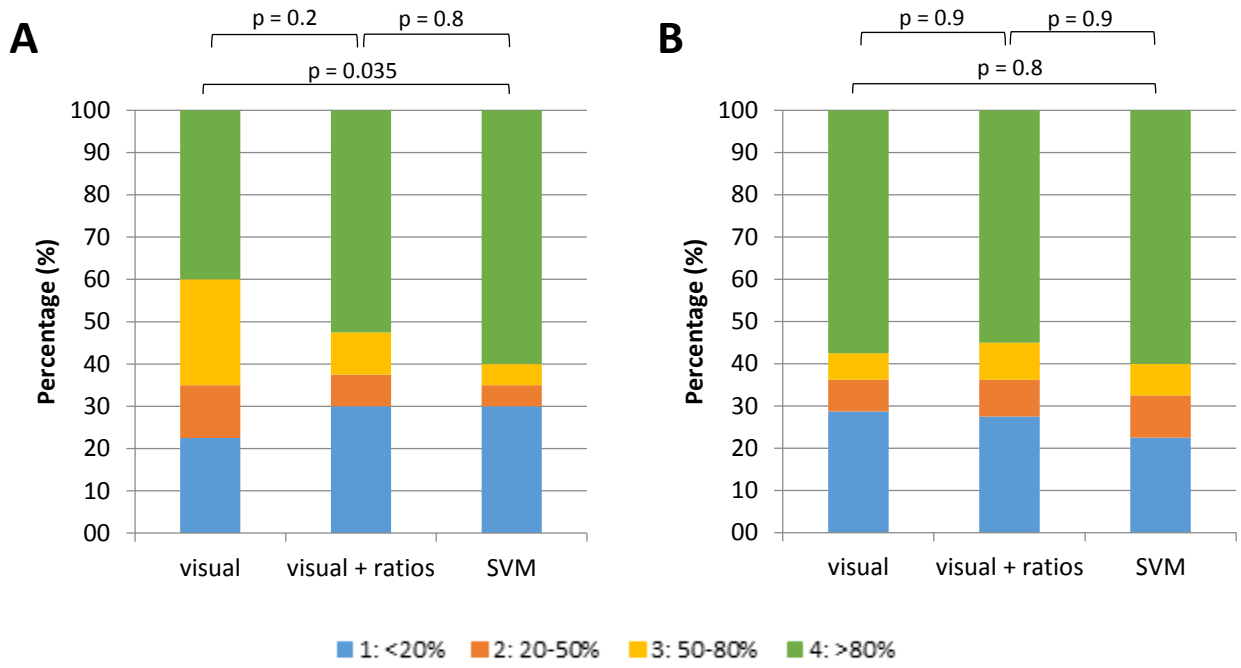


Figure 3: Frequencies in the scored chance of PD as determined by nuclear medicine physicians and the SVM model for the validation set of (A) Isala hospital and (B) Scheper hospital.

DISCUSSION

In this study, we derived a SVM model to discriminate PD from non-PD using uptake ratios, age and gender as input features that was able to determine whether input was characteristic for PD with high accuracy. Subsequent validation of the model using patients from the same center showed that a comparable prediction performance was obtained, indicating that the model is generalizable towards previously unseen data. Moreover, a similar prediction performance was found when the model was applied on data acquired in a different center with comparable acquisition and image processing methods. For both sites, performance was equivalent to that of nuclear medicine physicians who classified patients using standalone visual assessment of I-123 FP-CIT images and visual assessment combined with interpretation of semi-quantitative analysis results.

The high classification accuracy for the cross-validation found in this study is in line with several other studies that developed linear SVM models to evaluate SBRs derived from I-123 FP-CIT images. Palumbo et al. found an accuracy of 94.2% in a 5-fold cross-validation using 90 patients with SBRs and age as input features for discriminating PD from non-PD.[95] Prashanth et al. showed that their model was able to correctly discriminate healthy controls from early PD in 92.3% of cases (n=548) in a 10-fold cross-validation using only SBRs.[96] More recently, Taylor et al. compared different ML algorithms with a range of semi-quantification methods and showed that ML generated equal or higher mean accuracies than standalone semi-quantitative analysis, irrespective of the method used. Performing a 10-times repeated 10-fold cross-validation, they obtained a mean accuracy of 95% (n=657) and 89% (n=304) for their linear SVM models that were able to discriminate healthy controls from PD and Parkinsonian from non-Parkinsonian

patients, respectively.[91] Though high accuracies were found, none of these studies performed a subsequent validation using previously unseen data, nor did they assess whether derived models were valid across centers.

The prediction performance found for the so-called internal validation, usually performed via cross-validation, is not a guarantee for good generalizability as derived models usually perform better on the dataset used for training. Especially in small datasets, a decrease in prediction performance is often seen when the model is tested using an independent dataset. This can be due to overfitting, the lack of representative data used to train the model or the use of different imaging and processing protocols.[97–101] The similar prediction performance found of the derived model for both group B and C in comparison to group A indicate that overfitting did not occur. Acquisition, reconstruction and quantification methods are known to influence striatal uptake ratios irrespective of the presence of a neurodegenerative disease.[88, 102] Since a difference in the $Put_L/Caud_L$ index was found between only non-PD patients in group A and C, we can assume that the slight decrease in performance for group C in comparison to group B is due to a difference in patient population. Usage of the derived SVM model in different centers requires the use of comparable acquisition and image processing methods to ensure that findings are due to pathology and not due to variability in specifics used for acquisition, reconstruction and quantification. Furthermore, each center has to perform a pilot study to investigate how well the classifier can be generalized to one's own dataset and ideally, its performance is put into perspective by comparing it with that of nuclear medicine physicians.

This study has several limitations. First, a relatively small number of patients was used for training purposes that is not necessarily representative of all neurodegenerative and non-degenerative diseases that are evaluated using I-123 FP-CIT SPECT imaging. Other degenerative conditions that also show reduced tracer uptake are prone to be misclassified by the derived SVM model as I-123 FP-CIT SPECT imaging cannot reliably discriminate between neurodegenerative parkinsonian disorders.[103] Likewise, PD patients with scans without evidence for dopaminergic deficit (SWEDD) will presumably be incorrectly classified as non-PD. An increase in the number of other neurodegenerative diseases in group C could explain the lower median values of the ratios of non-PD patients in group A in comparison to group C and thus lower performance of the derived model as these patients are prone to misclassification. A larger patient population is needed to be able to identify neurodegenerative subtypes and SWEDD patients based on I-123 FP-CIT ratios, thereby deriving multiple SVM models as shown by Nicastro et al.[104] Second, information beyond the uptake ratios combined with gender and age was not considered for input data. As PD is a clinical diagnosis, the addition of parameters comprising the severity and progression of a patient's disease could have allowed better discrimination between PD and non-PD. This would require consistent assessment of patients suspected for PD, thereby using clinical rating scales such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale and Hoehn and Yahr scale.[105, 106] Furthermore, imaging features extracted from I-123 FP-CIT SPECT scans that comprise striatal shape have shown discriminative power in the identification of PD patients.[107–109] The addition of these features could have provided higher classification accuracies, but requires the development and validation of quantification methods to extract these features. In contrast to this, the current parameters used as input features are routinely collected in

clinical practice and therefore easily available. Third, different ML approaches including deep learning were not evaluated in discriminating PD from non-PD. Though these approaches could have been superior in performance to that of linear SVM, the derived model is relatively simple and transparent and can be easily exported and used in clinical practice. Finally, one needs to be aware that the derived model requires a consistent way of acquiring the different ratios in order to work properly. Automated approaches could overcome the variability associated with approaches that require manual steps. Nonetheless, the intra- and inter-reproducibility of the SVM output were excellent for the semi-quantitative method used.

The derived SVM model is an objective classification approach for identifying PD patients that has similar prediction performance as that of standard visual interpretation by expert nuclear medicine physicians. It can therefore facilitate clinical decision-making and diagnosis when used in clinical practice. Taylor et al. evaluated the impact of the addition of SVM-based interpretation of I-123 FP-CIT scans on clinical reporting. They found that consistency between reporters improved and that the model gave added confidence in terms of diagnostic confidence scores.[110] We can therefore assume that usage of our model in clinical practice can lead to less interpretation variation and more confident diagnosis of PD.

CONCLUSION

Development of a linear SVM model to interpret semi-quantitative analysis results from I-123 FP-CIT images allows high-accuracy detection of PD with similar classification accuracy as that of expert nuclear medicine physicians. The model is able to discriminate PD from non-PD and is feasible in centers using the same acquisition and image processing methods. The results of this study show that the use of the derived SVM model has great potential to be used in the diagnostic process of PD, thereby encouraging implementation of this SVM model in clinical practice.

CHAPTER 3

Development and validation of a machine learning-based approach for detection of myocardial ischemia and risk assessment of major adverse cardiac events

ABSTRACT

Aim: Refined assessment of patients at risk of coronary artery disease (CAD) using machine learning (ML) can provide a more individualized risk stratification approach to reduce the need for additional testing and to facilitate patient management. Our aim was to develop and validate a ML-based approach to detect myocardial ischemia and major adverse cardiac events (MACE).

Methods: We retrospectively included 6151 patients that underwent myocardial perfusion imaging (MPI) with SPECT. For a subset (n=2292), follow-up was available and patient outcome was reported. The total group was used to develop and validate a support vector machine (SVM) classifier with Gaussian kernel to discriminate patients with normal SPECT MPI scan from patients with abnormal scan. Another Gaussian SVM classifier to predict whether MACE occurs was built and evaluated using the subset. Both groups were divided into a training and validation dataset. Input features included various clinical parameters and coronary artery calcium (CAC) score for both models and for the MACE model, left ventricular ejection fraction (LVEF) and scan outcome were added. The validation sets were tested and used to assess the prediction performance of the derived models. The first model was further evaluated by comparing its performance to that of absent CAC indicative for normal MPI scan.

Results: Validation of the ischemia model led to a sensitivity, specificity, PPV, NPV of 89.7%, 31.8%, 29.0% and 90.8%, respectively. This was comparable to the performance of CAC scoring for predicting normal or abnormal scan ($p>0.14$). The MACE model was not generalizable beyond the training dataset as the specificity decreased from 16.5% to 3.1% ($p<0.001$) when using the validation dataset.

Conclusion: ML-based integration of clinical parameters and CAC score has comparable prediction performance as that of standalone CAC score for predicting myocardial ischemia. Further evaluation is needed for the MACE model to accurately assess whether patients are at risk of MACE.

INTRODUCTION

Myocardial perfusion imaging (MPI) with single photon emission computed tomography (SPECT) allows non-invasive evaluation of CAD. This widely available technique is used to identify and quantify the presence and extent of perfusion defects indicative of ischemia. SPECT MPI has a high prognostic value and is therefore important for risk stratification to optimize patient outcome.[47, 111, 112] It is recommended in individuals at intermediate risk of obstructive CAD.[47] The pre-test probability of CAD is often assessed by using the Diamond and Forrester classification, only requiring age, gender and symptom typicality to estimate the risk.[113] However, this classification is known to overestimate the prevalence of obstructive

CAD, thereby unnecessarily increasing the need for noninvasive or invasive testing while decreasing the yield of these tests.[114, 115]

More recently, the relationship between CAC and frequency of subsequent cardiac events has been well established which can be used to improve the selection of patients for MPI procedures and the assessment of overall patient risk. [116, 117] The addition of clinical risk factors and demographic features to the CAC-score have the potential to provide a more individualized risk assessment which may result in even better patient management and outcomes.[60] Machine learning is able to integrate all these different variables and has the potential to identify patients with myocardial ischemia or at risk of MACE. Hence, our aim was to develop and validate a machine learning-based approach to (1) integrate readily available clinical and functional parameters together with CAC score for better patient selection for SPECT MPI and (2) combine clinical and functional parameters with CAC score and SPECT outcome to predict MACE.

METHODS

Study population

We retrospectively included a total of 6151 patients who were suspected for CAD and underwent clinically indicated stress MPI with SPECT/CT with simultaneous assessment of coronary artery calcification CAC from January 2009 to March 2019. Multiple demographic and clinical features were available and derived from the medical records, including gender, age, weight and height. Furthermore, CAD risk factors, current medication use, blood pressure (BP), heart rate, level of creatinine in the blood and presence of left bundle branch block (LBBB) were collected. Patients did not have a known history of CAD, had no missing demographic or clinical features and CAC score and LVEF were available. All patients provided written informed consent for the use of their data for research purposes.

Patient preparation and image acquisition

Patients were instructed to refrain from caffeine or other methylxanthine-containing products for at least 24 hours prior to stress examination. Dipyridamole was discontinued for 48 hours before the test. Pharmacologic stress was induced by intravenous administration of adenosine (140 µg/kg/min for 6 minutes) or dobutamine (starting from 10 µg/kg/min, increased along 3-minute intervals to a maximum of 50 µg/kg/min until 85% of the predicted maximum heart rate was reached). At peak stress, 370 MBq (500 MBq for patients with a body weight >100 kg) Tc-99m tetrofosmin was injected intravenously. In case of abnormal stress perfusion, additional rest SPECT was performed on the same day using 740 MBq (1000 MBq for patient with a body weight >100 kg) Tc-99m tetrofosmin. Stress and rest imaging were performed 45-60 minutes after tracer injection and all patients underwent imaging in supine position with arms placed above their head. The time delay between stress and rest studies was >3 hours.

Patient scanned before May 2010 (n = 306) underwent imaging using a conventional dual-detector gamma camera (Ventri; GE Healthcare) equipped with a low energy, high resolution collimator. Images were acquired using a 20% symmetric window at 140 keV and 64x64 matrix and an elliptic orbit with step-and-shoot acquisition at 6° intervals over a 180° arc (45° anterior oblique to 45° left posterior oblique with 30 steps). Acquisition time was 12 minutes for the stress

images and 15 minutes for the rest images. Scans after May 2010 (n = 5845) were acquired using a cadmium zinc telluride (CZT)-based SPECT/CT camera (Discovery NM/CT 570c; GE Healthcare) equipped with 19 pinhole detectors each containing 32x32 pixelated (2.46x2.46 mm) CZT elements. Acquisition time was 5 minutes for the stress images and 4 minutes for the rest images.

Gated SPECT analysis was used to determine left ventricular volumes and ejection fraction and to assess motion wall abnormalities. After both the stress and rest imaging, an unenhanced low-dose CT scan during breath-hold was performed to provide an attenuation map of the chest for attenuation correction (LightSpeed VCT XT; GE Healthcare). The following scanning parameters were used: 5.0 mm slice thickness, 800 ms rotation time, pitch of 1.0, collimation 64x0.625 mm, tube voltage of 120 kV, tube current of 20 mA and an irradiation body length of 24.4 cm. Next, an unenhanced electrocardiographically-gated CT scan was obtained to calculate the CAC score. The CAC scan was triggered at 75% of the R-R interval by using a 2.5 mm slice thickness, gantry rotation time of 330 ms, tube voltage of 120 kV and a tube current of 125-250 mA, depending on patients' size.

Emission images and attenuation map data were processed with a dedicated reconstruction algorithm (Myovation; GE Healthcare) and displayed in the traditional short, long and horizontal axes. Along with this, 17-segments MPI bull's eye images were created that represented the percentage tracer uptake in the 17 myocardial segments with segmental uptake values normalized to the highest pixel value.[118] Post-processing of CAC-scans was performed using dedicated software (SmartScore, Advantage Windows 4.4, GE Healthcare) to calculate CAC score using standard Agatston criteria.[119]

SPECT MPI analysis and follow-up

Perfusion deficits on SPECT MPI scans were identified by expert nuclear physicians who interpreted the images and semi-quantitative results. The presence and size of irreversible and ischemic defects were reported together with the size of the defect (small, moderate, large). Based on these findings, patients were labelled as having an abnormal SPECT MPI scan or a normal scan.

For a subset of all included patients (n = 2292), a median follow-up of 36 months (range: 1-117 months) was available. The follow-up length was determined by assessing the interval between the examination date and the date of final consultation. Patient outcome including revascularization (both percutaneous coronary intervention and coronary bypass grafting), non-fatal myocardial infarction and cardiac death were previously assessed by reviewing patients' records, performing telephone interviews with the patients or by contacting patients' general practitioners. Based on this, patients were labelled as either having a positive outcome (patients in which MACE occurred) or negative outcome.

SVM development and validation

For both models, SVM with a Gaussian kernel was used as classification method. Continuous data were normalized such that the mean value was 0 and standard deviation was 1. All

procedures were performed using Matlab software (MATLAB and Statistics and Machine learning Toolbox Release 2018b, the Mathworks Inc.).

Ischemia model: the total population of 6151 patients was randomly divided into a training (n = 5540) and validation dataset (n = 611), thereby ensuring that the ratio of patients with abnormal and normal scans was comparable in both groups. Input features included age, gender, weight, height, CAD risk factors (current smoking, type II diabetes mellitus (DM), dyslipidemia and positive family history of CAD in first-degree relatives), medication use and type of medication (aspirin or clopidrogel, beta blockers, calcium channel blockers and statin), upper and lower BP, creatinine level, presence of LBBB and CAC score. The training group was used to build the SVM classifier and to perform hyperparameter optimization. A coarse grid search was performed, thereby evaluating regularization parameter and gamma values of 2^{-25}, 2^{-15}, 2^{-10}, …, 2^{25} after which a finer grid search was conducted to evaluate regularization parameter values of 2^{-10}, 2^{-9}, 2^{-8} …, 2^{10} and gamma values of 2^{15}, 2^{16}, 2^{17} …, 2^{25}. For each combination, a stratified 10-fold cross-validation was conducted after which sensitivity and specificity were determined. The combination of values that provided a sensitivity >90% and the highest specificity was selected to derive the final model.

MACE model: the subset of 2292 patients was randomly divided into a training (n = 1830) and validation dataset (n = 462), thereby ensuring that the ratio of patients with positive and negative outcome was equivalent in both groups. Input features included the features of the ischemia model aforementioned together with LVEF and SPECT MPI outcome (presence of irreversible defect and presence and size of ischemic defect). The training group was used to build the SVM classifier and to perform hyperparameter optimization. A coarse grid search was performed to select appropriate values of the regularization parameter and gamma that provided a sensitivity of 95%., thereby evaluating values between 2^{-25} and 2^{25}. Eventually, a finer grid search was conducted to evaluate regularization parameter values of 2^{-6}, 2^{-5}, 2^{-4} …, 2^{9} and gamma values of 2^{19}, 2^{20}, 2^{21} and 2^{22}. For each combination, a stratified 10-fold cross-validation was conducted after which sensitivity and specificity were determined. The combination of values that provided a sensitivity >95% and the highest specificity was selected to derive the final model.

Both models were tested using the validation sets, thereby assessing predicted classes for each patient. Prediction performance was determined by calculating sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Besides that, the area under the curve (AUC) of the derived model was assessed for the validation set.

Ischemia model vs. CAC

The prediction performance of the ischemia model was compared to that of CAC scoring alone. For the validation set, patients with a CAC score of zero were predicted as having a normal SPECT MPI scan and scores ≥ 1 as having an abnormal scan. Sensitivity, specificity, PPV and NPV were assessed for this method and compared to that of the SVM model.

Table 1: Baseline characteristics and scan outcomes of all patients who underwent clinically indicated SPECT MPI. Values are given for separate training and validation datasets used for the ischemia model and MACE model.

	Ischemia model			**MACE model**		
Characteristic	training (n = 5540)	validation (n = 611)	p-value (χ^2 / t-test)	training (n=1830)	validation (n=462)	p-value (χ^2 / t-test)
Age (years)	61.7 ± 11.4	62.1 ± 11.7	0.40	60.8 ± 11.7	61.0 ± 11.2	0.74
Male gender	44.0%	41.7%	0.31	42.7%	47.4%	0.08
Body weight (kg)	81.6 ± 15.7	82.2 ± 15.4	0.34	81.9 ± 16.0	82.2 ± 14.6	0.77
Height (cm)	172.7 ± 9.7	172.3 ± 9.5	0.35	172.3 ± 9.8	172.3 ± 9.8	0.94
Current smoking	14.1%	12.9%	0.45	14.7%	14.9%	0.96
Hypertension	56.6%	61.2%	0.032	60.4%	58.0%	0.37
DM type II	12.7%	10.8%	0.21	12.1%	15.8%	0.043
Dyslipidemia	38.4%	37.8%	0.80	40.9%	40.9%	>0.99
Family history	55.0%	51.4%	0.10	53.5%	56.5%	0.27
Medication use	83.8%	86.8%	0.07	83.4%	83.8%	0.92
LBBB	3.2%	2.5%	0.38	3.4%	3.5%	>0.99
Systolic BP	138.8 ± 20.7	140.0 ± 21.3	0.20	139.8 ± 21.0	137.9 ± 20.2	0.08
Diastolic BP	84.6 ± 12.2	84.9 ± 12.5	0.53	84.8 ± 12.6	83.6 ± 12.8	0.07
Heart rate	71.1 ± 12.4	71.1 ± 13.1	0.94	71.1 ± 12.4	70.6 ± 12.0	0.42
Creatinine level	80.6 ± 38.1	79.4 ± 17.8	0.15	77.2 ± 28.0	76.2 ± 17.3	0.30
CAC	297 ± 676	336 ± 701	0.19	273 ± 612	354 ± 703	0.023
LEVF	62.0 ± 9.4	62.0 ± 9.6	0.90	62.2 ± 9.2	62.1 ± 10.6	0.97
Normal MPI scan	76.5%	76.3%	0.95	75.1%	71.9%	0.17
MACE	-	-	-	6.7%	6.9%	0.96

LBBB, left bundle branch block; BP, blood pressure; CAC, coronary artery calcification; LVEF, left ventricular ejection fraction; MPI, myocardial perfusion imaging; MACE, major adverse cardiac risk events. Values are presented as mean ± SD or percentages.

Statistical analysis

Statistical analysis was performed using R Studio software (RStudio: Integrated Development for R. Version 1.1.442; RStudio, Inc.). To assess differences in patient characteristics between training and validation datasets, the t-test or χ^2-test were performed. A χ^2-test was also used to compare prediction performances of the derived models between the training and validation groups. Differences in sensitivities and specificities of the ischemia model and CAC scoring were evaluated using the McNemar's test and differences in PPV and NPV were assessed using a weighted generalized score statistic.[120] For all statistical tests, a significance level at 0.05 was used.

RESULTS

The baseline characteristics and clinical parameters of all included patients are summarized in Table 1. For both models, the majority of all parameters was comparable between the training and validation set. The validation set used for the ischemia model contained a higher number of patients with hypertension than the training set ($p = 0.032$). Besides that, we observed that more patients with DM type II were present in the validation set of the MACE model ($p = 0.043$) and that CAC scores were significantly higher in this group in comparison to the training group ($p = 0.023$).

Model development and validation

For the 10-fold cross-validation, the derived ischemia model (regularization parameter of 2^{-7} and gamma value of 2^{17}) had a sensitivity of 91.5% and specificity of 27.5% with PPV and NPV of 28.0% and 91.3%, respectively. The model showed goodg generalization towards unseen data as all prediction performance parameters assessed for the validation dataset were comparable to that found for the training set ($p > 0.06$) as seen in Table 2. The derived model lead to an AUC of 0.71 for the validation set. For the derived MACE model (regularization parameter of 2^3 and gamma value of 2^{20}), a 10-fold cross-validation resulted in a sensitivity, specificity, PPV and NPV of 95.9%, 16.5%, 7.6% and 98.3%, respectively. Testing the model using the validation dataset, a substantial decrease in specificity (specificity = 3.3%, $p < 0.001$) was found in comparison to the training set. The other parameters were comparable ($p > 0.6$) between training and validation datasets as shown in Table 2. For the MACE model, an AUC of 0.77 was found for the validation set.

Table 2: Prediction performance of the ischemia and MACE models for both training and validation set.

	Ischemia			**MACE**		
Parameter	Training	Validation	p-value (χ^2-test)	Training	Validation	p-value (χ^2-test)
Sensitivity	91.5%	89.7%	0.6	95.9%	96.8%	>0.99
Specificity	27.5%	31.8%	0.06	16.5%	3.3%	<0.001
PPV	28.0%	29.0%	0.7	7.6%	6.7%	0.6
NPV	91.3%	90.8%	0.95	98.3%	93.3%	0.7

PPV, positive predictive value; NPV, negative predictive value

Ischemia model vs. CAC

Prediction performance of the ischemia model and CAC scoring using the validation dataset is shown in Figure 1. For the ischemia model, a sensitivity, specificity, PPV and NPV of 89.7%, 31.8%, 29.0% and 90.8% was obtained, respectively. Using a CAC-score of zero as indicative for normal SPECT, a comparable sensitivity of 84.1% ($p = 0.14$), specificity of 37.1% ($p = 0.17$), PPV of 29.4% ($p = 0.7$) and NPV of 88.3% ($p = 0.3$) was found.

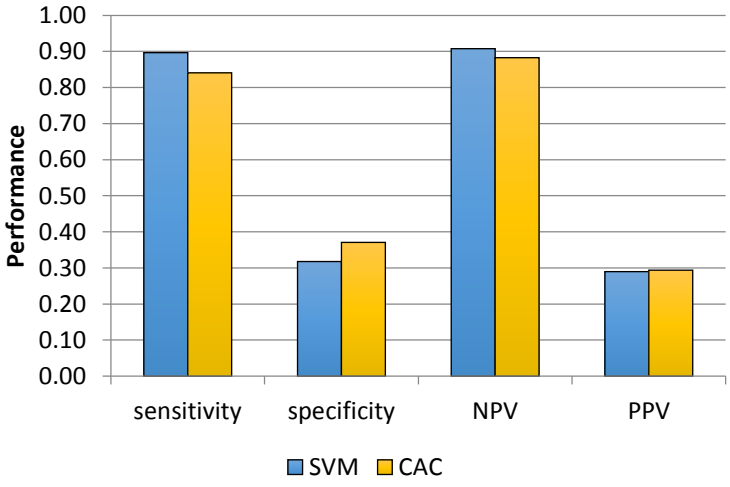


Figure 1: Prediction performance of the SVM model and CAC score to discriminate patients with an abnormal SPECT MPI scan from patients with a normal scan.

DISCUSSION

In this study, we developed two SVM models that integrated readily available clinical parameters with imaging outcomes to provide a more individualized risk assessment of myocardial ischemia and MACE. The derived ischemia model showed good generalization towards previously unseen data, in contrast to the MACE model for which a considerable increase in false positives was observed for the validation dataset. ML-based integration of clinical parameters and CAC score had comparable prediction performance as that of standalone CAC score for predicting whether abnormalities are present on SPECT MPI scans.

To our knowledge, no other studies have assessed the ability of machine learning to integrate clinical parameters and CAC score to predict the presence of ischemia on MPI SPECT and subsequently compared the prediction performance of the derived model to that of standalone CAC scoring. Juarez et al. evaluated the feasibility and performance of ML in identifying patients with (1) ischemia and (2) elevated risk of MACE as determined by Nitrogen-13 ammonia positron emission tomography (PET) imaging. Using only accessible clinical and functional parameters as input, they evaluated a boosted ensemble ML algorithm in a stratified 10-fold cross-validation and reached an AUC of 0.72 and 0.71, respectively. Testing the model on a previously unseen dataset resulted in an AUC of 0.75 for both models. The derived models showed superior performance to that of readily available models that are recommended in the

European Society of Cardiology guidelines.[121] The AUCs found in this study are comparable to the AUC of 0.72 and 0.77 found in our study for the prediction of ischemia and occurrence of MACE, respectively. Juarez et al. used the myocardial flow reserve (MFR) for labelling purposes, whereas we used nuclear medicine physician's interpretation of both images and semi-quantitative data to determine whether ischemia was present and follow-up results to assess whether MACE occurred. MFR is known for its incremental diagnostic value in detecting CAD when used as an adjunct to PET MPI and its significant prognostic value in predicting MACE.[122–126] Although our approach is time-consuming, it is a more accurate resemblance of how ischemia is detected in clinical practice and it uses true events for labelling purposes of the MACE model. In this way, models can be built that capture a more realistic relationship between the input and output, thereby resulting in better prediction performances. As PET is superior to SPECT in the detection of ischemia[127–129], the use of PET data could yield even better prediction performances if further input features are kept comparable.

Several other studies have evaluated the use of machine learning to predict MACE using both clinical parameters and quantitative SPECT MPI image features. Alonso et al. derived a SVM model to predict cardiac death that reached an AUC of 0.83.[130] Arsanjani et al. used a boosted ensemble ML algorithm to predict revascularization events and achieved an AUC of 0.81, which was found comparable and superior to experienced expert readers and better than standalone measures of perfusion derived from SPECT MPI.[131] Likewise, Betancur et al. found an AUC of 0.81 for a boosted ensemble ML model to predict MACE that was superior to existing visual and automated perfusion assessment.[132] Though a comparable AUC of 0.77 was found in this study, comparison is difficult as each study defined its own input and output parameters and none of the aforementioned studies performed a validation using previously unseen data. All three studies showed that (semi-)quantitative MPI data substantially contributed to the model, indicating that the addition of these parameters could potentially enhance the performance of the current MACE model. This is in line with previous findings that semi-quantitative parameters including summed stress score, summed rest score and summed difference score objectively reflect the extent and severity of perfusion defects, thereby complementing conventional image interpretation and having prognostic value in the prediction of MACE.[118, 133–135]

For both models, we aimed for a high sensitivity that apparently required a high gamma value as observed when performing the grid search. High gamma values can indicate a degree of overfitting.[72, 136] The MACE model is thereby more prone to overfitting as the training dataset contained a relative small proportion of positive examples (patients in which MACE occurred) in comparison to the ischemia model. This can explain why we observe a good generalizability for the ischemia model, whereas we see a substantial decrease in specificity of the MACE model for the validation set. It is not certain whether this decrease is due to the selected gamma value as not all patient characteristics were comparable between the training and validation sets. We observed that the validation set contained a higher amount of patients with DM type II and that the mean CAC score was higher. DM type II is associated with faster CAD progression and is therefore an independent predictor of adverse outcomes.[137–139] Likewise, an increase in the occurrence of MACE is observed with increasing CAC scores.[116, 140, 141] It is possible that there was an insufficient amount of patients with these characteristics in the training set to learn from that

subsequently lead to an increase in the false positive rate. Further subanalyses are therefore needed in order to determine whether the derived model is suitable for patients with DM type II and higher CAC scores or that retraining is needed.

We found that a ML-based approach for predicting whether abnormalities are present on SPECT MPI is comparable to standalone CAC scoring in which a score of zero is indicative for a normal scan. For clinical use, the latter would be preferred as assessing a single parameter is less time-consuming. The addition of demographic and clinical parameters was expected to further increase the ability to accurately predict abnormal SPECT MPI, but this was not observed in this study. Though input features were selected based on expert opinion, not all the features may contribute to the prediction of ischemia. Liu et al. found that a model using only a few predictors outperformed a whole set of variables in predicting MACE[142], indicating the need of selecting an appropriate combination of input features that provides the highest prediction performance. Multiple methods can be used for this purpose.[143, 144] Juarez et al. determined the information gain of a set of predictor variables for myocardial ischemia and excluded those that documented no gain.[121] Several of these excluded features, including BMI, sex and smoking, are used as input features in our derived SVM model, suggesting that the current combination of features provide a suboptimal prediction performance. Therefore, future work should focus on feature selection methods to assess whether used input features impede prediction performance. In this way, a model can be derived that is expected to demonstrate the added value of clinical parameters besides CAC scoring, thereby potentially providing a more individualized risk stratification.

This study has several limitations. First, the patient population used for the development and validation of the ML models is rather large, but data was assessed from one center and follow-up was not extensive for every patient. It is not known how well the derived model will perform in different centers and patient populations, nor whether the derived model will be suitable for future patient populations. Second, the MACE model was not compared to current recommended risk scores due to its poor generalizability and therefore, the clinical value of this model was not assessed. Finally, SVM is one of many ML algorithms that is able to map the relationship between a defined input and output. Different ML approaches including deep learning were not evaluated in predicting ischemia and MACE, but could yield better prediction performances. The use of ML has the potential to provide an automated risk estimate that combines multiple variables to identify patients with ischemia or at risk of MACE, though further research of the current derived SVM models is needed before clinical implementation can be considered. Further refinement of the models could provide a more accurate assessment of patients that require noninvasive SPECT or more invasive evaluation and treatment.

CONCLUSION

ML-based integration of clinical parameters and CAC score has comparable prediction performance as that of standalone CAC score for predicting whether abnormalities will be present on SPECT MPI. Therefore, further evaluation is needed to determine the discriminative power of these clinical features in predicting myocardial ischemia. The derived SVM model for risk assessment of MACE can predict the occurrence of MACE with high sensitivity, but a significant reduction in the number of false positives is needed before implementation in clinical practice can be considered.

CHAPTER 4

Recommendations and future perspectives

The studies presented in this thesis show that the use of machine learning (ML) can integrate multiple variables to make individualized predictions for diagnosis and prognosis purposes. Although not all developed support vector machine models presented in this thesis currently have the potential to be clinically implemented, the studies show that ML is applicable in different fields. It is therefore plausible that ML also has impact in applications beyond the detection of Parkinson's disease, myocardial ischemia and major adverse cardiac events. These not only include individualized models based on information extracted from nuclear images and patient's medical health records to facilitate patient management, but potential applications lie also in the field of image reconstruction and processing as de-noising, segmentation and registration.[145, 146] Though the use of ML seems promising, ML tools have limitations and several aspects have to be taken into account considering the development, validation and clinical implementation process.

The availability of data is one of the main challenges in the development of ML models, since a large amount of training data with sufficient quality is required to provide reliable results.[147] Careful labelling is needed as inaccuracies will limit prediction performance of the model. Therefore, the involvement of experts is often required as well as substantial amounts of time and effort to ensure data accurately reflects clinical reality.[13, 148] Data sharing between multiple institutions could give access to a large number of proven cases that can be used for both development and validation, but methods for quality control of shared data and images and support for data transfer and storage without impeding data privacy are yet to be developed. Regarding the generalizability of ML models, criteria should be established that define in what patient population and under which circumstances a derived model is valid.[149] One example comprises the image acquisition and reconstruction settings as these are known to influence features extracted from images, thereby impeding generalizability across centers. Further standardization of these settings can potentially facilitate a more wide-spread use of ML.[13, 150]

Advances in the field of natural language processing (NLP) and radiomics would further facilitate the development of ML-based applications. NLP systems are able to generate structured information from unstructured free text as found in electronic medical records, thereby enabling automatic identification and extraction of clinical information.[151, 152] Radiomics involves the high-throughput extraction of quantitative imaging features that characterize a particular regions of interest (ROI), thereby capturing information related to pathophysiology. These features are hand-crafted and comprise different aspects of a segmented ROI including intensity, shape and texture.[150, 153] Both NLP and radiomics techniques have the potential to provide input parameters that can be easily obtained once established methods are available. Regarding medical images, the use of deep learning can overcome the need for defining and selecting features as well as the segmentation step that is required to extract the features of a ROI. 154] In deep learning, the algorithm can automatically and adaptively learn feature representation from raw

data by summarizing and transforming clusters of pixels in the image.[5, 148, 155] It has therefore the potential to play a key role in the discovery of new imaging biomarkers that allows us to identify and monitor diseases.[146]

Considering the implementation of ML models in clinical practice, interpretability and explainability of a model is essential for its acceptance and usability in the clinical setting. This comprises the ability to provide an understanding of input parameters, model parameters, the algorithm used and the reasons behind a model's prediction results.[156, 157] The latter is increasingly addressed through the assessment of the most relevant features and data and model visualization using visualization techniques such as nomograms and salient maps. Understanding in how predictions are generated by the model results in a gain of trust of its users and it allows the possibility to correct errors.[158, 159]

From a regulatory perspective, ML models need to be certified before large-scale deployment to ensure safe use of these tools in clinical practice.[148] The Food and Drug Administration and medical devices certification systems are likely to play a key role in approving ML applications, but rules about datasets, transparency and verification procedures are currently lacking.[160] Existing laws regarding regulated medical devices are difficult to apply to ML algorithms as the models will evolve over time when more data is processed and learned from. Periodic testing over specific time intervals could potentially ensure that prediction performance is improving consistently and will not decline.[5] Other issues may arise from the use of patient data for training purposes and questions regarding liability.[5, 149, 160] This calls for the need to further develop existing regulatory frameworks to establish an appropriate regulation of ML-based applications in healthcare.

Integration of ML tools in clinical practice will likely evolve the role of radiologists and nuclear medicine physician. It is plausible that the physician will contribute to the training and refinement of ML models as they can add expert knowledge and experience to the tools they will then use. Therefore, a basic understanding of the methods and concepts of ML should be part of their training.[145] As ML models are restricted to one well-defined task, it is assumed that ML will take over tasks that are more routine and standardized, while assisting the physician in tasks that require more context.[13] For each application, the optimal clinical workflow has to be identified and the effect of use of models in real-life clinical situations has to be assessed. An integrated ML component in the clinical setting is thought to improve quality of care by reducing human error, increasing diagnostic certainty, supporting workload management and facilitating personalized medicine.[6, 7, 148, 149]

REFERENCES

1. Langs G, Röhrich S, Hofmanninger J, Prayer F. Machine learning: from radiomics to discovery and routine. Radiologe 2018;58:S1–S6.

2. Lee J, Jun S, Cho Y, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging : General Overview. Korean J Radiol 2017;18:570–584.

3. Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. J Appl Biomed 2013;11:47–58.

4. Choi JY. Radiomics and Deep Learning in Clinical Imaging: What Should We Do? Nucl Med Mol Imaging (2010) 2018;52:89–90.

5. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–510.

6. Fazal MI, Patel ME, Tye J, Gupta Y. The past, present and future role of artificial intelligence in imaging. Eur J Radiol 2018;105:246–250.

7. Rahim MK, Kim SE, So H, Kim HJ, Cheon GJ, Lee ES, et al. Recent Trends in PET Image Interpretations Using Volumetric and Texture-based Quantification Methods in Nuclear Oncology. Nucl Med Mol Imaging (2010) 2014;48:1–15.

8. Choi H. Deep Learning in Nuclear Medicine and Molecular Imaging: Current Perspectives and Future Directions. Nucl Med Mol Imaging 2018;52:109–118.

9. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. RadioGraphics 2017;37:505–515.

10. Giger ML. Machine Learning in Medical Imaging. J Am Coll Radiol 2018;15:512–520.

11. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018;2:1–10.

12. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. Am J Roentgenol 2017;208:1–7.

13. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20:e262–e273.

14. Suzuki K. Pixel-Based Machine Learning in Medical Imaging. Int J Biomed Imaging 2012;2012:1–18.

15. Wang S, Summers RM. Machine learning and radiology. Med Image Anal 2012;16:933–951.

16. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology 2018;00:1–11.

17. Lin M-C, Iqbal U, Li Y-C. AI in Medicine: Big Data Remains a Challenge. Comput Methods Programs Biomed 2018;164:A1.

18. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys 2019;46:e1–e36.

19. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–510.

20. Aerts HJWL. Data science in radiology: A path forward. Clin Cancer Res 2018;24:532–534.

21. Lee J-G, Jun S, Cho Y-W, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol 2017;18:570–584.

22. Sajda P. Machine Learning for Detection and Diagnosis of Disease. Annu Rev Biomed Eng 2006;8:537–565.

23. Tysnes OB, Storstein A. Epidemiology of Parkinson's disease. J Neural Transm 2017;124:901–905.

24. Ascherio A, Schwarzschild MA. The epidemiology of Parkinson's disease: risk factors and prevention. Lancet Neurol 2016;15:1257–1272.

25. Findley LJ. The economic impact of Parkinson s disease. Park Relat Disord 2007;13:8–12.

26. Reeve A, Simcox E, Turnbull D. Ageing and Parkinson's disease: Why is advancing age the biggest risk factor ? Ageing Res Rev 2014;14:19–30.

27. Maresova P, Klimova B, Novotny M, Kuca K. Alzheimer's and Parkinson' s Diseases: Expected Economic Impact on Europe - A Call for a Uniform European Strategy. J Alzheimer's Dis 2016;54:1123–1133.

28. Poewe W, Seppi K, Tanner CM, Halliday GM, Volkmann J, Schrag A, et al. Parkinson disease. Dis Prim 2017;3:1–21.

29. Beitz JM. Parkinson's disease: a review. Front Biosci 2014;S6:65–74.

30. Liu Z, Liu F-T, Zuo C-T, Koprich JB, Wang J. Update on Molecular Imaging in Parkinson ' s Disease. Neurosci Bull 2018;34:330–340.

31. National Institute for Health and Care Excellence. Parkinson's disease in adults: diagnosis and management. NICE Guidel. No. 71

32. Schreglmann SR, Bhatia KP, Stamelou M. Advances in the Clinical Differential Diagnosis of Parkinson's Disease. Parkinsons Dis 2017;132:79–127.

33. Postuma RB, Berg D, Stern M, Poewe W, Marek K, Litvan I. MDS Clinical Diagnostic Criteria for Parkinson 's Disease. Mov Disord 2015;30:1591–1599.

34. Rizzo G, Copetti M, Arcuti S, Martino D. Accuracy of clinical diagnosis of Parkinson disease A systematic review and meta-analysis. Am Acad Neurol 2016;86:1–11.

35. Saeed U, Compagnone J, Aviv RI, Strafella AP, Black SE, Lang AE, et al. Imaging biomarkers in Parkinson's disease and Parkinsonian syndromes : current and emerging concepts. Transl Neurodegener 2017;6:1–25.

36. Wu X, Cai H, Ge R, Li L, Jia Z. Recent Progress of Imaging Agents for Parkinson's Disease. Curr Neuropharmacol 2014;12:551–563.

37. Bloem BR, van Laar T, Keus SHJ. Multidisciplinaire richtlijn Ziekte van Parkinson 2006-2010. Alphen aan den Rijn: Van Zuiden Communications; 2010.

38. Tripathi M, Kumar A, Bal C. Neuroimaging in Parkinsonian Disorders. Neurol India 2018;66:68–78.

39. Brooks DJ. Imaging Approaches to Parkinson Disease. J Nucl Med 2010;51:596–610.

40. Calabria FF, Calabria E, Gangemi V, Cascini GL. Current status and future challenges of brain imaging with 18F- DOPA PET fot movement disorders. Hell J Nucl Med 2016;19:33–41.

41. Park E. A New Era of Clinical Dopamine Transporter Imaging Using 123I-FP-CIT. J Nucl Med Technol 2012;40:222–228.

42. Djang DSW, Janssen MJR, Bohnen N, Booij J, Henderson TA, Herholz K, et al. SNM Practice Guideline for Dopamine Transporter Imaging with 123I-Ioflupane SPECT 1.0. J Nucl Med 2012;53:154–163.

43. Ogawa T, Fujii S, Kuya K, Kitao S-I, Shinohara Y, Ishibashi M, et al. Role of Neuroimaging on Differentiation of Parkinson's Disease and Its Related Diseases. Yonago Acta Med 2018;61:145–155.

44. Nichols M, Townsend N, Scarborough P, Rayner M. Cardiovascular disease in Europe 2014 : epidemiological update. 2014;1–10.

45. Hansson GK. Inflammation, Atherosclerosis, and Coronary Artery Disease. N Engl J Med 2005;352:1685–1695.

46. Menees DS, Bates ER. Evaluation of patients with suspected coronary artery disease. Coron Artery Dis 2010;21:386–390.

47. Montalescot G, Sechtem U, Achenbach S, Andreotti F. 2013 ESC guidelines on the management of stable coronary artery disease. Eur Heart J 2013;34:2949–3003.

48. Klein L, Gheorghiade M. Coronary artery disease and prevention of heart failure. Med Clin North Am 2004;88:1209–1235.

49. Cornelissen A, Jinnouchi H, Sakamoto A, Torii S, Kuntz S, Guo L, et al. Evaluation and Management of the Vulnerable Plaque. Curr Cardiovasc Risk Rep 2019;13:1–11.

50. Cassar A, Holmes DR, Rihal CS, Gersh BJ, Gps N, Qbujfou F. Chronic Coronary Artery Disease : Diagnosis and Management. Mayo Clin Proc 2009;84:1130–1146.

51. Mordi IR, Badar AA, John Irving R, Weir-McCall JR, Houston JG, Lang CC. Efficacy of noninvasive cardiac imaging tests in diagnosis and management of stable coronary artery disease. Vasc Health Risk Manag 2017;13:427–437.

52. Gamble JHP, Scott G, Ormerod JOM, Frenneaux MP. Pathophysiology of coronary artery disease: The case for multiparametric imaging. Expert Rev Cardiovasc Ther 2009;7:299–310.

53. Alexanderson-Rosas E, Guinto-Nishimura GY, Cruz-Mendoza JR, Oropeza-Aguilar M, De La Fuente-Mancera JC, Barrero-Mier AF, et al. Current and future trends in multimodality imaging of coronary artery disease. Expert Rev Cardiovasc Ther 2015;13:715–731.

54. Heo R, Nakazato R, Kalra D, Min JK. Noninvasive imaging in coronary artery disease. Semin Nucl Med 2014;44:398–409.

55. Pakkal M, Raj V, McCann GP. Non-invasive imaging in coronary artery disease including anatomical and functional evaluation of ischaemia and viability assessment. Br. J. Radiol. 2011;84(3):S280-295

56. Iwasaki K. Myocardial ischemia is a key factor in the management of stable coronary artery disease. World J Cardiol 2014;6:130–139.

57. Lee JH, Han D, Danad I, Hartaigh B, Lin FY, Min JK. Multimodality imaging in coronary artery disease: Focus on computed tomography. J Cardiovasc Ultrasound 2016;24:7–17.

58. Qiu L, Tan H, Cheng D, Shi H. The incremental clinical value of cardiac hybrid SPECT/CTA imaging in coronary artery disease. Nucl Med Commun 2018; 39(6):469-478.

59. Weininger M, Renker M, Rowe GW, Abro JA, Costello P, Schoepf UJ. Integrative computed tomographic imaging of coronary artery disease. Expert Rev Cardiovasc Ther 2011;9:27–43.

60. Zellweger MJ. Risk stratification in coronary artery disease: a patient-tailored approach over the ischaemic cascade. Swiss Med Wkly 2019;149:1–9.

61. Collet C, Onuma Y, Miyazaki Y, Morel M-A, Serruys PW. Integration of non-invasive functional assessments with anatomical risk stratification in complex coronary artery disease: the non-invasive functional SYNTAX score. Cardiovasc Diagn Ther 2017;7:151–158.

62. Marsland S. Machine Learning: an Algorithmic Perspective, 2nd ed. Boca Raton: Taylor & Francis Group; 2015.

63. Simeone O. A Brief Introduction to Machine Learning for Engineers. Found Trends Signal Process 2018;12:200–431.

64. Bishop CM. Pattern Recognition and Machine Learning. Singapore: Springer Science and Business Media; 2006.

65. Hastie T. Overview of Supervised Learning. In: Hastie T, Tibshirani R, Friedman F (eds) Elem. Stat. Learn., 2nd ed. Springer Science and Business Media, New York, 2009; 9–41

66. Ghahramani Z. Unsupervised Learning. In: Bousquet O (ed) Mach. Learn. Springer, Berlin, 2004; 72–73

67. Khan SG, Herrmann G, Lewis FL, Pipe T, Melhuish C. Reinforcement learning and optimal adaptive control : An overview and implementation examples. Annu Rev Control 2012;36:42–59.

68. Neumann J Von. Model selection and overfitting. Nat Methods 2016;13:703–704.

69. Wong T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit 2015;48:2839–2846.

70. Baumann K. Cross-validation as the objective function for variable-selection techniques. Trends Anal Chem 2003;22:395–406.

71. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surv 2010;4:40–79.

72. Ben-Hur A, Weston J. A user's guide to support vector machines. In: Carugo O, Eisenhaber F (eds) Data Min. Tech. life Sci. Humana Press, New Delhi, 2010; 223–239

73. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: A review of classification and combining techniques. Artif Intell Rev 2006;26:159–190.

74. Wernick MN, Yang Y, Brankov JG, Yourganov G. Machine Learning in Medical Imaging. IEEE Signal Process Mag 2018;27:25–38.

75. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn 1995;20:273–297.

76. Mallinson H, Gammerman A. Imputation Using Support Vector Machines. 2003. http://www.cs.york.ac.uk/euredit/

77. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–1567.

78. Ben-Hur A, Weston J. A user's guide to support vector machines. Methods Mol Biol 2010;609:223–239.

79. Seeböck P. Deep Learning in Medical Image Analysis. Technical University of Vienna

80. Booij J, Knol RJJ. SPECT imaging of the dopaminergic system in (premotor) Parkinson's disease. Park Relat Disord 2007;13:425–428.

81. Catafau AM, Tolosa E. Impact of Dopamine Transporter SPECT Using 123 I-Ioflupane on Diagnosis and Management of Patients With Clinically Uncertain Parkinsonian Syndromes. Mov Disord 2004;19:1175–1182.

82. Marshall VL, Reininger CB, Marquardt M, Patterson J, Hadley DM, Oertel WH, et al. Parkinson's Disease is Overdiagnosed Clinically at Baseline in Diagnostically Uncertain Cases: A 3-Year European Multicenter Study with Repeat [123 I] FP-CIT SPECT. Mov Disord 2009;24:500–508.

83. Bairactaris C, Demakopoulos N, Tripsianis G, Sioka C, Farmakiotis D. Impact of dopamine transporter single photon emission computed tomography imaging using I-123 ioflupane on diagnoses of patients with parkinsonian syndromes. J Clin Neurosci 2009;16:246–252.

84. Darcourt J, Booij J, Tatsch K, Varrone A, Vander Borght T, Kapucu ÖL, et al. EANM procedure guidelines for brain neurotransmission SPECT using 123 I-labelled dopamine transporter ligands, version 2. Eur J Nucl Med Mol Imaging 2009;37:443–450.

85. Badiavas K, Molyvda E, Iakovou I, Tsolaki M, Psarrakos K, Karatzas N. SPECT imaging evaluation in movement disorders: Far beyond visual assessment. Eur J Nucl Med Mol Imaging 2011;38:764–773.

86. Booij J, Dubroff J, Pryma D, Yu J, Agarwal R, Lakhani P, et al. Diagnostic Performance of the Visual Reading of 123 I-Ioflupane SPECT Images With or Without Quantification in Patients With Movement Disorders or Dementia . J Nucl Med 2017;58:1821–1826.

87. Ueda J, Yoshimura H, Shimizu K, Hino M, Kohara N. Combined visual and semi-quantitative assessment of 123 I-FP-CIT SPECT for the diagnosis of dopaminergic neurodegenerative diseases. Neurol Sci 2017;38:1187–1191.

88. Tossici-bolt L, Dickson JC, Sera T, Nijs R De, Bagnara MC, Jonsson C, et al. Calibration of gamma camera systems for a multicentre European 123 I-FP-CIT SPECT normal database. Eur J Nucl Mol Imaging 2011;38:1529–1540.

89. Tossici-bolt L, Dickson JC, Sera T, Booij J, Asenbaun-nan S, Bagnara MC, et al. [123 I] FP-CIT ENC-DAT normal database : the impact of the reconstruction and quantification methods. EJNMMI Phys 2017;4:1–16.

90. Cascianelli S, Scialpi M, Amici S, Forini N, Minestrini M, Fravolini M, et al. Role of Artificial Intelligence Techniques (Automatic Classifiers) in Molecular Imaging Modalities in Neurodegenerative Diseases. Curr Alzheimer Res 2017;14:198–207.

91. Taylor JC, Fenner JW. Comparison of machine learning and semi-quantification algorithms for (I123)FP-CIT classification: the beginning of the end for semi-quantification? EJNMMI Phys 2017;4:1–20.

92. Taylor JC, Romanowski C, Lorenz, E, et al.Computer-aided diagnosis for (123 I)FP-CIT imaging: impact on clinical reporting. EJNMMI Res 2018;8:1–10.

93. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine Learning in Health Care: A Critical Appraisal of Challenges and Opportunities. J Electron Heal Data Methods 2019;7:1–4.

94. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv Large Margin Classif 1999;61-74.

95. Palumbo B, Fravolini M, Buresta T, Pompili T. Diagnostic Accuracy of Parkinson Disease by Support Vector Machine (SVM) Analysis of 123I-FP-CIT Brain SPECT Data. Medicine (Baltimore) 2014;93:e228.

96. Prashanth R, Dutta Roy S, Mandal PK, Ghosh S. Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging. Expert Syst Appl 2014;41:3333–3342.

97. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: A clinical example. J Clin Epidemiol 2003;56:826–832.

98. Huf W, Kalcher K, Boubela RN, Rath G, Vecsei A, Filzmoser P, et al. On the generalizability of resting-state fMRI machine learning classifiers. Front. Hum. Neurosci. 8:

99. de Bruijne M. Machine learning approaches in medical image analysis: From detection to diagnosis. Med Image Anal 2016;33:94–97.

100. Therrien R, Doyle S. Role of training data variability on classifier performance and generalizability. Proc. SPIE 10581, Medical Imaging 2018: Digital Pathology, 1058109

101. Withney H, Ji Y, Li H, Edwards A, Papaioannou J. Effect of diversity of patient population and acquisition systems on the use of radiomics and machine learning for classification of 2397 breast lesions. Proc. SPIE 10950, Medical Imaging 2019: Computer-Aided Diagnosis, 109501A

102. Tossici-bolt L, Dickson JC, Sera T, Booij J, Asenbaun-nan S, Bagnara MC, et al. [123 I] FP-CIT ENC-DAT normal database: the impact of the reconstruction and quantification methods. EJNMMI Phys 2017;4:1–16.

103. Kagi G, Bhatia K, Tolosa E. The role of DAT-SPECT in movement disorders. J Neurol Neurosurg Psychiatry 2010;81:5–12.

104. Nicastro N, Garibotto V, Poncet A, Badoud S, Burkhard PR. Establishing On-Site Reference Values for 123 I-FP-CIT SPECT (DaTSCAN®) Using a Cohort of Individuals with Non-Degenerative Conditions. Mol Imaging Biol 2016;18:302–312.

105. Goetz CG, Poewe W, Rascol O, Sampaio C, Stebbins GT, Counsell C, et al. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations. Mov Disord 2004;19:1020–1028.

106. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov Disord 2008;23:2129–2170.

107. Segovia F, Górriz JM, Ramírez J, Martínez-Murcia FJ. Automatic Separarion of Parkisonian Patients and Control Subjects Based on the Striatal Morphology. 2017;10337:345–352.

108. Oliveira FPM, Faria DB, Costa DC, Castelo-Branco M, Tavares JMRS. Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson's disease based on [123 I]FP-CIT SPECT images. Eur J Nucl Med Mol Imaging 2017;45:1052–1062.

109. Bhalchandra NA, Prashanth R, Roy SD, Noronha S. Early detection of Parkinson's disease through shape based features from 123I-Ioflupane SPECT imaging. Proc. Int. Symp. Biomed. Imaging. IEEE Computer Society 2015; 963–966

110. Taylor JC, Romanowski C, Lorenz E, Lo C, Bandmann O, Fenner J. Computer-aided diagnosis for (123 I) FP-CIT imaging: impact on clinical reporting. EJNMMI Res 2018;8:1–10.

111. Gaemperli O, Saraste A, Knuuti J. Cardiac Hybrid imaging. Eur Heart J 2012;13:51–60.

112. Flotats A, Knuuti J, Gutberlet M, Marcassa C, Bengel FM, Kaufmann PA, et al. Hybrid cardiac imaging: SPECT/CT and PET/CT. A joint position statement by the European Association of Nuclear Medicine (EANM), the European Society of Cardiac Radiology (ESCR) and the European Council of Nuclear Cardiology (ECNC). Eur J Nucl Med Mol Imaging 2011;38:201–212.

113. Diamond GA, Forrester JS. Analysis of Probability as an Aid in the Clinical Diagnosis of Coronary-Artery Disease. N Engl J Med 2010;300:1350–1358.

114. Genders TSS, Steyerberg EW, Alkadhi H, Leschka S, Desbiolles L, Nieman K, et al. A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension. Eur Heart J 2011;32:1316–1330.

115. Bittencourt MS, Hulten E, Polonsky TS, Hoffman U, Nasir K, Abbara S, et al. European society of cardiology-recommended coronary artery disease consortium pretest probability scores more accurately predict obstructive coronary disease and cardiovascular events than the diamond and forrester score. Circulation 2016;134:201–211.

116. Engbers EM, Timmer JR, Ottervanger JP, Mouden M, Knollema S, Jager PL. Prognostic value of coronary artery calcium scoring in addition to single-photon emission computed tomographic myocardial perfusion imaging in symptomatic patients. Circ Cardiovasc Imaging 2016;9:1–9.

117. Mouden M, Timmer JR, Reiffers S, Oostdijk AHJ, Knollema S, Ottervanger JP, et al. Coronary Artery Calcium Scoring to Exclude Flow-Limiting Coronary Artery Disease in Symptomatic Stable Patients at Low or Intermediate Risk. Radiology 2013;269:77–83.

118. Berman D, Abidoy A, Kang X, Hayes S, Friedman J. Prognostic validation of a 17-segment score derived from a 20-segment score for myocardial perfusion SPECT interpretation. J Nucl Cardiol 2004;11:414–423.

119. Agatson A, Janowitz W, Hildner F, Zusmer N. Quantification of coronary artery calcium using ultrafast computed tomography. J Am Coll Cardiol 1990;15:827–832.

120. Kosinski AS. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. Stat Med 2013;32:964–977.

121. 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.

122. Fiechter M, Ghadri JR, Gebhard C, Fuchs TA, Pazhenkottil AP, Nkoulou RN, et al. Diagnostic Value of 13N-Ammonia Myocardial Perfusion PET: Added Value of Myocardial Flow Reserve. J Nucl Med 2012;53:1230–1234.

123. Berti V, Sciagrà R, Neglia D, Pietilä M, Scholte AJ, Nekolla S, et al. Segmental quantitative myocardial perfusion with PET for the detection of significant coronary artery disease in patients with stable angina. Eur J Nucl Med Mol Imaging 2016;43:1522–1529.

124. Herzog BA, Husmann L, Valenta I, Gaemperli O, Siegrist PT, Tay FM, et al. Long-Term Prognostic Value of 13N-Ammonia Myocardial Perfusion Positron Emission Tomography. J Am Coll Cardiol 2009;54:150–156.

125. Ziadi MC, Dekemp RA, Williams KA, Guo A, Chow BJW, Renaud JM, et al. Impaired myocardial flow reserve on rubidium-82 positron emission tomography imaging predicts adverse outcomes in patients assessed for myocardial ischemia. J Am Coll Cardiol 2011;58:740–748.

126. Ziadi MC. Myocardial flow reserve (MFR) with positron emission tomography (PET)/computed tomography (CT): clinical impact in diagnosis and prognosis. Cardiovasc Diagn Ther 2017;7:206–218.

127. Bateman T, Heller G, McGhie A, et al. Diagnostic accuracy of rest/stress ECG-gated Rb-82 myocardial perfusion PET: Comparison with ECG-gated Tc-99m sestamibi SEPCT. J Nucl Cardiol 2006;13:24–33.

128. Bengel F, Higuchi T, Javadi M, et al. Cardiac Positron Emission Tomography. J Am Coll Cardiol 2009;54:1–15.

129. Arumugam P, Tout D, Tonge C. Myocardial perfusion scintigraphy using rubidium-82 position emission tomography. Br Med Bull 2013;107:87–100.

130. Haro Alonso D, Wernick MN, Yang Y, Germano G, Berman DS, Slomka P. Prediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning. J Nucl Cardiol 2018;1–9.

131. Arsanjani R, Dey D, Khachatryan T, Shalev A, Hayes SW, Fish M, et al. Prediction of revascularization after myocardial perfusion SPECT by machine learning in a large population. J Nucl Cardiol 2015;22:877–884.

132. 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. JACC Cardiovasc Imaging 2018;11:1000–1009.

133. Sharir T, Slomka P. Dual-isotope myocardial perfusion SPECT imaging: Past, present, and future. J Nucl Cardiol 2017;25:2024–2028.

134. Leslie WD, Tully SA, Yogendran MS, Ward LM, Nour KA, Metge CJ. Prognostic value of automated quantification of 99mTc-sestamibi myocardial perfusion imaging. J Nucl Med 2005;46:204–211.

135. Berman DS, Kang X, Van Train KF, Lewin HC. Comparative prognostic value of automatic quantitative analysis versus semiquantitative visual analysis of exercise myocardial perfusion single-photon emission computed tomography. J Am Coll Cardiol 1998;32:1987–1995.

136. Varewyck M, Martens JP. A practical approach to model selection for support vector machines with a Gaussian kernel. IEEE Trans Syst Man, Cybern Part B Cybern 2011;41:330–340.

137. Blanke P, Naoum C, Ahmadi A, et al. Long-term prognostic utility of coronary CT angiography in stable patients with diabetes mellitus. JACC Cardiovasc Imaging 2016;9:1280–1288.

138. Nakanishi R, Ceponiene I, Osawa K, et al. Plaque progression assessed by a novel semiquantitative plaque software on coronary computed tomography angiography between diabetes and non-diabetes patients: a propensity-scoring matching study. Atherosclerosis 2016;255:73–79.

139. Bayturan O, Tuzcu E, Uno K, et al. Comparison of rates of progression of coronary atherosclerosis in patients with diabetes mellitus versus those with the metabolic syndrome. J Am Coll Cardiol 2010;105:1735–1739.

140. Chang SM, Nabi F, Xu J, Peterson LE, Achari A, Pratt CM, et al. The Coronary Artery Calcium Score and Stress Myocardial Perfusion Imaging Provide Independent and Complementary Prediction of Cardiac Risk. J Am Coll Cardiol 2009;54:1872–1882.

141. Hou ZH, Lu B, Gao Y, Jiang SL, Wang Y, Li W, et al. Prognostic value of coronary CT angiography and calcium score for major adverse cardiac events in outpatients. JACC Cardiovasc Imaging 2012;5:990–999.

142. Liu N, Koh ZX, Goh J, Lin Z, Haaland B, Ting BP, et al. Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. BMC Med Inform Decis Mak 2014;14:1–9.

143. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70–79.

144. Chandrashekar G, Sahin F. A survey on feature selection methods. J Comput Electr Eng 2014;40:16–28.

145. Visvikis D, Cheze Le Rest C, Jaouen V, Hatt M. Artificial intelligence, machine (deep) learning and radio(geno)mics: definitions and nuclear medicine imaging applications. Eur J Nucl Med Mol Imaging

146. Choi H. Deep Learning in Nuclear Medicine and Molecular Imaging: Current Perspectives and Future Directions. Nucl Med Mol Imaging (2010) 2017;52:109–118.

147. Uribe CF, Mathotaarachchi S, Gaudet V, Smith KC, Rosa-Neto P, Bénard F, et al. Machine Learning in Nuclear Medicine: Part 1—Introduction. J Nucl Med 2019;60:451–458.

148. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719–731.

149. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. J Am Coll Radiol 2018;15:504–508.

150. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749–762.

151. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: A systematic review. Radiology 2016;279:329–343.

152. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med 2018;131:129–133.

153. Ibrahim A, Vallières M, Woodruff H, Primakov S, Beheshti M, Keek S, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. Semin Nucl Med

154. Acharya UR, Hagiwara Y, Sudarshan VK, Chan WY, Ng KH. Towards precision medicine: from quantitative imaging to radiomics. J Zhejiang Univ Sci B 2018;19:6–24.

155. Aerts HJWL. Data science in radiology: A path forward. Clin Cancer Res 2018;24:532–534.

156. Bibal A, Frénay B. Interpretability of machine learning models and representations: An introduction. ESANN 2016 - 24th Eur. Symp. Artif. Neural Networks. i6doc.com publication, 2016; 77–82

157. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. Proc. - 2018 IEEE 5th Int. Conf. Data Sci. Adv. Anal. DSAA 2018. Institute of Electrical and Electronics Engineers Inc. 2019; 80–89

158. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput. Appl. 2019 https://doi.org/10.1007/s00521-019-04051-w

159. Biran N-Join O, Cotton N-Join C. Explanation and Justification in Machine Learning: A Survey. IJCAI 2017. Workshop on Explainable Artificial Intelligence (XAI), 8–13

160. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. Am J Roentgenol 2017;208:754–760.

161. Lavalaye J, Booij J, Reneman L, Habraken J. Effect of age and gender on dopamine transporter imaging with [123I] FP-CIT SPET in healthy volunteers. Eur J Nucl Med 2000;27:867–869.

162. Varrone A, Dickson JC, Tossici-Bolt L, Sera T, Asenbaum S, Booij J, et al. European multicentre database of healthy controls for [123 I]FP- CIT SPECT (ENC-DAT): Age-related effects, gender differences and evaluation of different methods of analysis. Eur J Nucl Med Mol Imaging 2013;40:213–227.

163. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn 2002;46:389–422.

164. Mladenić D, Brank J, Grobelnik M, Milic-Frayling N. Feature selection using linear classifier weights. Association for Computing Machinery (ACM), 2004; 234-241

APPENDIX A

SVM model development and characteristics for detection of Parkinson's Disease

For the development of a machine learning (ML) model, several steps have to be taken to derive a proper working model. First, a specific regression or classification problem has to be defined that ML is going to address. Subsequently, output and input parameters can be selected as well as a suitable algorithm. After data collection, the data is split into a training and validation dataset after which the data is preprocessed. A method has to be selected to perform hyperparameter optimization, thereby defining an evaluation measure in order to select the most suitable hyperparameter values. Finally, a set of prediction performance parameters have to be selected and assessed to obtain an estimate of the derived model's performance. Aforementioned aspects are separately addressed for the development of a ML-based approach for interpretation of I-123 FP-CIT SPECT scans to detect Parkinson's disease (PD).

Algorithm and feature selection

The aim of our study was to discriminate PD from diagnoses other than PD (non-PD) based on I-123 FP-CIT SPECT scans, thereby defining a binary classification problem. The output parameter comprised the clinical diagnosis retrieved from a patient's medical record**:** an output value of 1 was given to PD patients, while non-PD patient had an output value of 0. Input parameters included I-123 FP-CIT striatal uptake ratios that are routinely assessed in clinical practice. Furthermore, we added age and gender as these are known to influence the ratios.[161, 162] Support vector machine (SVM) was selected as ML algorithm as this algorithm is known for its good generalizability properties.[64, 75] Based on visual plots of the ratios for both PD and non-PD patients, the groups were assumed to be linearly separable and a linear kernel was therefore chosen.

Data split and preprocessing steps

After all data was collected, 70% was used to create the training dataset and the remaining 30% formed the validation dataset. We ensured that both groups contained an equivalent amount of PD and non-PD patients. To ensure each feature is treated with equal importance by the SVM model, all input features, except for gender, were normalized according to:

$$z = \frac{x_i - \mu}{\sigma} \qquad (1)$$

Using the mean (μ) and standard deviation (σ) of the training dataset, the z-score (z) can be returned of a particular feature of one patient (x_i). Mean and standard deviation for each feature of the training dataset are found in Table 1. Gender was converted into two input features, 'male' and 'female'. Input feature 'male' had a value of 1 and 'female' a value of 0 when the patient was male (vice versa for a female patient). Furthermore, features of all patients were provided to the model in one specific order.

Table 1: Mean and SD of all continuous features of the training data and final model weights for all features

	Mean ± SD	Weight
male	-	-0.013
female	-	0.013
age	67.0 ± 10.9	-0.268
str. L	3.10 ± 0.90	-0.162
str. R	3.02 ± 0.83	-0.227
caud. L	3.58 ± 0.93	-0.055
caud. R	3.47 ± 0.84	0.146
put. L	2.67 ±0.92	-0.234
put. R	2.63 ± 0.87	-0.287
put/caud. L	0.74 ± 0.11	-0.454
put/caud. R	0.75 ± 0.11	-0.372

str, striatum; caud, caudate nucleus; put, putamen; L, left; R, right

Hyperparameter optimization

A linear SVM model comprises one hyperparameter, the regularization term C, of which an appropriate value had to be selected. Therefore, a grid search was performed on a logarithmic scale, thereby evaluating values of 2^{-6}, 2^{-5}, …, 2^{7}. For each value, a stratified, 10-fold cross-validation was conducted that was repeated 10 times. That is, the training dataset was partitioned into 10 folds with in each fold a similar class distribution to that in the training dataset. By repeating the cross-validation multiple times, thereby having different random partitions of the training dataset into the 10 folds, an estimation of the true performance of the model can be derived. As classification accuracy in unbalanced datasets tends to undervalue how well a model is able to correctly classify unseen examples, the mean F1-score was used as evaluation measure as this measure balances precision and recall. The value providing the highest F1-score was eventually selected to be used in the final derivation of the model. Mean and standard deviation of the F1-score for each value of C in the grid search is illustrated in Figure 1.

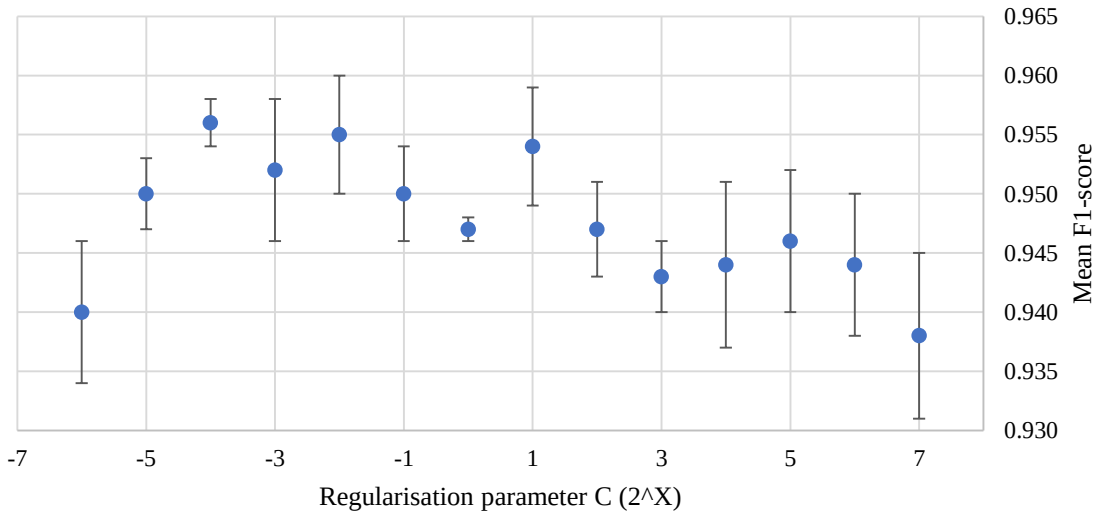


Figure 1: Mean F1-score and standard deviation for all values of C conducted in the grid search, assessed by performing a stratified, 10-times repeated 10-fold cross-validation.

Final model parameters and predictions

The eventual model was derived using a regularization parameter of 2^{-4}, thereby obtaining the fitted linear coefficients or weights for each feature (found in Table 1) and a bias term (found in Table 2). The weight magnitude of the different input features reflect the importance of the features.[163, 164] Therefore, we can determine which features are relevant to the model. The top three most important features in this model includes both left and right putamen/caudates indices as well as the specific binding ratio of the right putamen. Gender is the least relevant feature to the model to classify new examples.

In case a new prediction has to be made for an unseen example, a vector of all aforementioned features (x) is first normalized according to Formula (1). Using the weights (θ, in vector form) and the bias term (β), the classification score (S) can be obtained:

$$S = x \cdot \theta + \beta \qquad (2)$$

The classification score reflects the distance from x to the decision boundary and is either positive or negative. A positive score indicates that x is classified as PD, while a negative score indicates that the model classifies the example as being non-PD. To derive a probability that an example belongs to PD, a sigmoid function is derived using the inbuilt function <fitSVMPosterior> in Matlab (MATLAB and Statistics and Machine learning Toolbox Release 2018b, the Mathworks Inc.). Parameters comprising the scale (a) and intercept (c) of the sigmoid function are determined (found in Table 2) using this function after which the probability of PD (P) can be assessed:

$$P = 1 - \frac{1}{1+e^{-a(S-c)}} \qquad (3)$$

In conclusion, an interpretable model was derived that subsequently was validated as described in Chapter 2. This Appendix provides all means to make new predictions for previously unseen examples, but caution should be taken when applied in a different center in which the proposed ML model is not validated (yet).

Table 2: Derived values for the bias term and scale and intercept parameters of the sigmoid function.

parameter	value
β	0.592
a	-2.349
c	-0.017

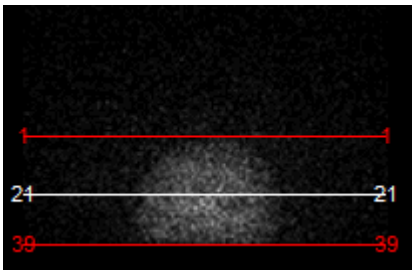
APPENDIX B

Protocol reconstructie en uitwerking van I-123 FP-CIT scans

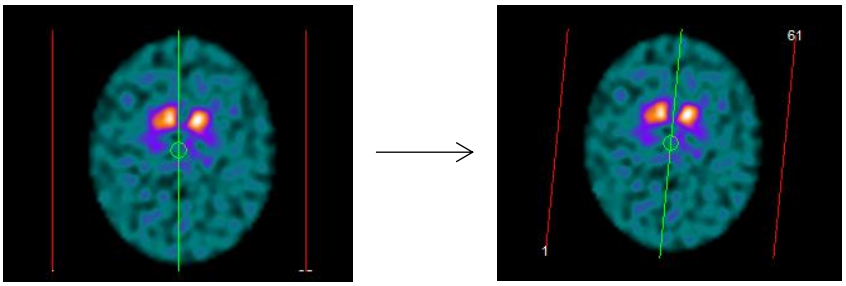
Reconstructie

1. Selecteer gehele patiënten map en klik op All applications >> Brain >> Brain SPECT
2. Selecteer uit het lijstje de FPCIT (Dataset Name) en vul in bij Data Usage Confirmation (wanneer gevraagd):
    a. Scan Group: -EARLY
    b. Modify usage >> Emission
3. Ga naar Rec/Ref >> Recon/Reformat/Mask

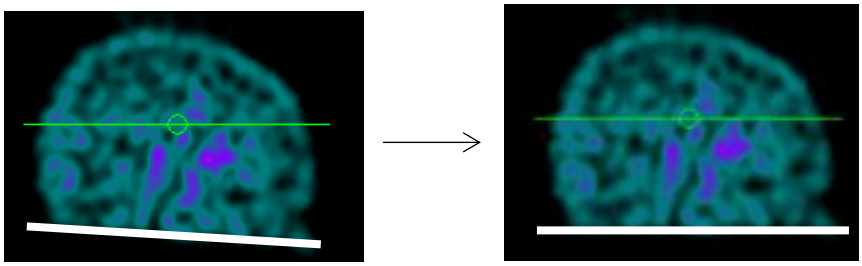
4. Pas de rode lijnen in het cine-beeld dusdanig aan dat de hersenen goed zijn omsloten. Zet de witte lijn op een slice waarin sagittaal het striatum goed zichtbaar is.



5. Zorg ervoor dat de activiteit in het striatum goed zichtbaar is in het overview en oriënteer de hersenen op de juiste manier.
    a. Sagittal Limits: groene lijn in het midden van linker- en rechter striatum. Eventueel de as draaien zodat het striatum recht komt te liggen (te controleren in transversale overview).



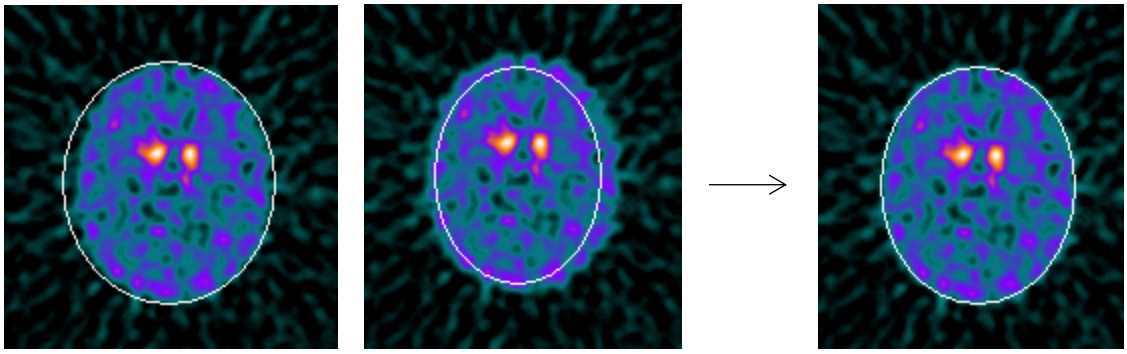
b. Transversal Limits: zet de groene lijn dusdanig neer dat in het transversale overview te zien is dat de activiteit in het striatum toeneemt en weer afneemt. Eventueel de as draaien zodat de as evenwijdig loopt met de acanthomeatale lijn (wit).



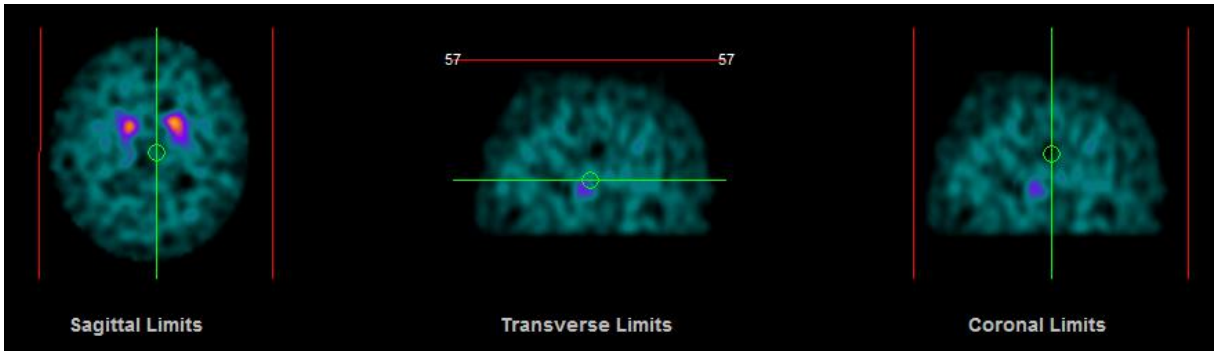
c. Coronal limits: zet de groene lijn dusdanig neer dat in het coronale overview te zien is dat de activiteit in het striatum toeneemt en weer afneemt.

6. Ga naar SPECT options en pas de instellingen eventueel aan. **Het is erg belangrijk dat deze goed staan**!
    a. Reconstruction options
        i. Recontruction type: FBP
    b. Correction options
        i. Scatter correction: Off
        ii. Attenuation correction
            1. Method: Change order 0
            2. Coefficient 0.11
            3. Threshold: 15 (standaard)
    c. Filters
        i. Pre-filter: Butterworth
        ii. Critical frequency: 0.65
        iii. Power: 10
        iv. RampFilter: Quantitative

7. Om de transversale sneden plaatst zich een region. Zet in je cine-view links je groene lijn dusdanig neer dat je het striatum in beeld krijgt en pas de threshold aan zodat de region de contouren van het hoofd volgt. Controleer de fit met behulp van de de groene lijn naast de transversale snede, scroll door het hoofd heen en kijk of de region het hoofd goed omsluit in elke slice. Wanneer correct: proceed. Nu verschijnt de attenuatie-gecorrigeerde uitwerking van de SPECT.



8. Scroll door je cine-view en controleer of het hoofd compleet binnen de rode lijnen vallen in je sagittal, transversal en coronal limits. Pas ze eventueel aan.



9. Ga naar Review

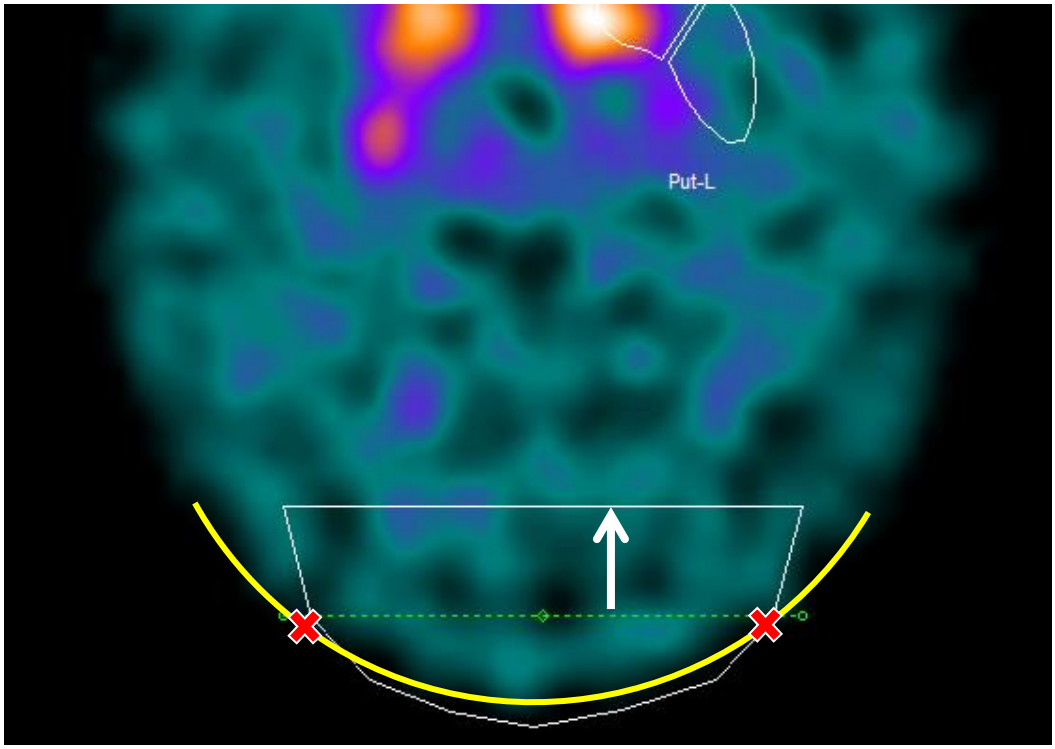
10. Kies bij plane >> transversal, je krijgt nu een overzicht met transversale beelden
11. Gebruik de zoom functie om de hersenen beeldvullend in beeld te brengen (zoom ca. 2.5). Zorg dat de kleuren setting op GE COL staat.
12. Maak een screencap via EF-screencap
13. Save de plaatjes via File >> save and exit

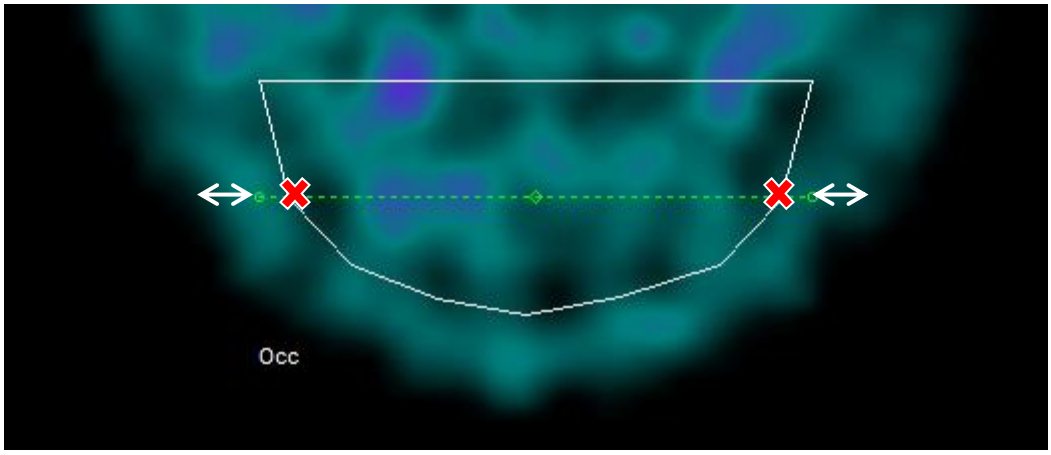
Kwantificatie

14. Selecteer de gesavede studie en werk deze uit met het protocol: DatSCAN 3r21
15. Selecteer de snede waarin het striatum het duidelijkste zichtbaar is qua volume en qua activiteit >> proceed
16. Positioneer de regions. Door de regions aan te klikken wordt deze geactiveerd en kan deze verplaatst en gekanteld worden.
    a. Occipital region: het is belangrijk dat deze region op eenzelfde hoogte komt te liggen voor elke patiënt. Hiervoor moeten twee stappen worden uitgevoerd:

i. de region dusdanig verplaatsen naar beneden totdat de punten waar de witte en groene lijnen elkaar kruizen (rode kruizen) op de grens van het hoofd komen te liggen (gele lijn).

ii. De platte kant van de region is nu de hoogte waar de groene lijn terecht moet komen (zie witte pijl). Verschuif de region naar deze hoogte en zorg dat er evenveel afstand is tussen de groene bolletjes en de grens van het hoofd. **Kantel de region niet!**

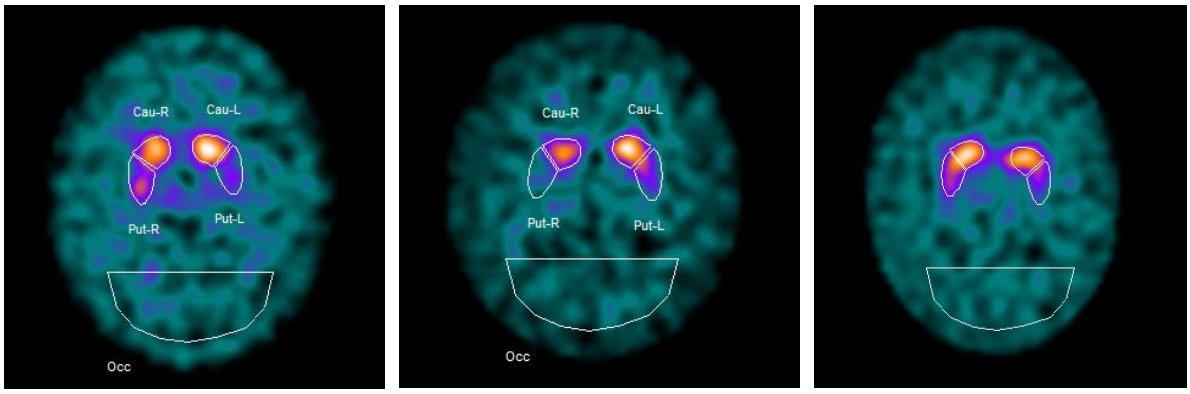




b. Striatum regions: begin bij het putamen en plaats deze waar (je denkt dat) je de staart kan zien. Zet de andere region om de nucleus caudates en zorg ervoor dat

ze zo goed mogelijk op elkaar aansluiten (zie voorbeelden hieronder). **Ze mogen niet overlappen!**

c. Proceed. De regions worden nu gekopieerd naar rechts. Herhaal stap b voor de rechterkant.



17. Calculate results. Controleer alle slices en pas waar nodig je putamen en nucleus caudates regions aan. De occipitaal region moet voor alle slices gelijk blijven. Zonodig herhalen via 'Start over'.

18. Controleer de acquisitie en reconstructie parameters. Wanneer het Butterworth filter niet de juiste waarde aangeeft is er iets mis gegaan in de reconstructie.
    a. Dataset Name: FBPAC_Transversal Obl
    b. Acquisition Zoom: 1.28
    c. Slice thickness: 3.45 mm
    d. Filter: Butterworth 0.65 / 10
    e. Attenuation corrected: YES
    f. Attenuation Corr Type: CHANG
    g. Attenuation Coeff: 0.11

19. Maak een screencap via EF-screencap
20. Save de studie via File >> save and exit

21. Stuur alles m.u.v. de result series naar SECTRA

APPENDIX C

Accepted abstracts

Submitted and accepted abstracts for the European Association of Nuclear Medicine (page 47) and Dutch Society of Nuclear Medicine and (page 48).

Machine learning-based interpretation of I-123-FP-CIT scans allows high-accuracy detection of Parkinson's Disease

M Dotinga[1,3], JD van Dijk[1], BN Vendel[1], CH Slump[3], JA van Dalen[2]

Departments of [1]Nuclear Medicine and [2]Medical Physics, Isala, Zwolle, The Netherlands and [3]MIRA: Institute for Biomedical Technology and Technical Medicine, University of Twente, Enschede, The Netherlands

Introduction: Dopamine transporter SPECT imaging with I-123-FP-CIT allows for visualisation of dysfunction of the dopaminergic system, which is characteristic of Parkinson's Disease (PD). Interpretation of scans based on visual assessment and semi-quantitative analysis imposes limitations as the latter requires a site-specific reference database that is often not available. Our aim was to develop a machine learning (ML)-based approach for interpretation of I-123-FP-CIT scans and determine its added value in clinical practice.

Methods: We retrospectively included a consecutive cohort of 130 patients that underwent I-123-FP-CIT SPECT imaging (Discovery D670, GE Healthcare) and had a clinically confirmed diagnosis. Patients were labelled as either having PD or a diagnosis other than PD (non-PD) and divided into a training set (58 PD, 32 non-PD) and validation set (25 PD, 15 non-PD) using stratified random sampling. The training set was used to build a linear support vector machine (SVM) classifier to discriminate PD from non-PD using I-123-FP-CIT striatal uptake ratios, age and gender as input features. Ratios were obtained by means of semi-quantitative analysis (Xeleris 4.0, GE Healthcare) and comprised specific binding in striatum, caudate nucleus and putamen as well as a putamen/caudate index for both left and right hemisphere. A stratified, 10-times repeated 10-fold cross-validation was conducted to perform model optimization using mean accuracy and F_1-score as evaluation measures. Subsequently, the derived SVM model was tested on the validation set. I-123-FP-CIT scans and corresponding ratios of the validation set were scored as either PD or non-PD by two expert nuclear medicine physicians following European guidelines. Overread from a third expert was performed in case of disagreement. Next, their prediction performance was compared to that of the SVM model.

Results: The highest mean prediction accuracy and F_1-score as found by cross-validation were 94.3% and 0.956, respectively. Testing the derived SVM model on the validation set, an accuracy of 95.0%, sensitivity of 96.0% and specificity of 93.3% were obtained. Prediction performance did not differ from visual assessment of PD, obtaining an equivalent accuracy, sensitivity and specificity of 95.0%, 96.0% and 93.3% ($p > 0.99$), respectively.

Conclusion: ML-based interpretation of I-123-FP-CIT scans results in accurate discrimination of PD from non-PD identical to standard visual assessment, thereby encouraging implementation of this SVM model as diagnostic aid in clinical practice.

Machine learning-based interpretation of I-123-FP-CIT striatal uptake ratios allows high-accuracy detection of Parkinson's Disease

M Dotinga[1,3], JD van Dijk[1], BN Vendel[1], CH Slump[3], JA van Dalen[2]

Departments of [1]Nuclear Medicine and [2]Medical Physics, Isala, Zwolle, The Netherlands and [3]MIRA: Institute for Biomedical Technology and Technical Medicine, University of Twente, Enschede, The Netherlands

Introduction: Dopamine transporter SPECT imaging (I-123-FP-CIT) allows for visualisation of dysfunction of the dopaminergic system, which is characteristic of Parkinson's Disease (PD). Interpretation based on visual assessment and semi-quantitative analysis imposes limitations as the latter requires a site-specific reference database that is often not available. Our aim was to develop and validate a machine learning (ML)-based interpretation of these I-123-FP-CIT scans to discriminate Parkinson's Disease (PD) from non-PD.
Methods: We retrospectively included 130 consecutive patients that underwent I-123-FP-CIT SPECT imaging (Discovery D670, GE Healthcare) and had a clinically confirmed diagnosis. Patients were labelled as either having PD or non-PD and divided into a training set (58 PD, 32 non-PD) and validation set (25 PD, 15 non-PD) using stratified random sampling. The training set was used to build a linear support vector machine (SVM) classifier using I-123-FP-CIT striatal uptake ratios, age and gender as input features. Subsequently, the derived SVM model was tested on the validation set and its prediction performance was compared to that of an expert nuclear medicine physician who scored the I-123-FP-CIT scans as either PD or non-PD.
Results: Accuracy, sensitivity and specificity of visual assessment of PD were 87.5%, 92.0% and 80.0%, respectively. A comparable prediction performance was found when using the SVM model, obtaining an accuracy of 95.0%, sensitivity of 96.0% and specificity of 93.3% ($p>0.24$).
Conclusion: ML-based interpretation of I-123-FP-CIT scans results in accurate discrimination of PD from non-PD similar to standard visual assessment, thereby encouraging implementation of this SVM model as diagnostic aid in clinical practice.