Deventer ziekenhuis

RAM ROBOTICS AND MECHATRONICS

AN AUTOMATIC SEGMENTATION METHOD AND PREDICTION MODEL FOR SKIN PRICK TEST RESULTS

M.S.M. (Marieke) Geessinck

MSC ASSIGNMENT

Committee:
dr. ir. F. van der Heijden
R.F.M. van Doremalen, MSc
dr. ir. B.J.F. van Beijnum
dr. J. Faber

July 2021

047RaM2021
Robotics and Mechatronics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | TECHMED CENTRE

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Summary

The Skin Prick Test (SPT) is the first step in allergy diagnostics and it is a widely used tool for over many decades. The inter-observer variability and measurement errors of this test lead to a lack of objectivity and reproducibility, which makes inter-institutional comparison of test results challenging. Besides, the current prediction model of SPT results lead to a relatively high amount of unnecessary follow-up diagnosis with additional costs. This research project aims to automate and decentralize the SPT and improve the patient outcome prediction. By doing so, a more accurate, quantitative, objective and reproducible allergy diagnosis can be made.

In order to automate and decentralize the SPT reading process, a deep learning network is proposed. This network extracts the wheal areas from digital photographs taken from the patient's forearm. The photographs are pre-processed and a deep Residual U-net (ResUnet) is trained on the training data, annotated by the author and verified by three assistants of the outpatient's clinic. The results show a dice similarity coefficient of 0.77, an intersection-over-union of 0.55 and an accuracy of 0.91. The method shows a similar accuracy with a higher precision, compared to computer vision based algorithms presented in literature.

To fulfill the aim of an improved prediction model, the predictive accuracy of the SPT is improved considering two strategies. Initially, the SPT measurements error are reduced by development of a semi-automatic algorithm that extracts wheal characteristics from digitized papersheets of the SPT results. Secondly, multiple clinical predictors are incorporated into a more complex predictive model. Five different prediction models have been evaluated: the cutoff based model as currently used in clinic and four machine learning approaches; Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB) and Logistic Regression (LR). The results show that both an improved wheal size determination and a more complex predictive model incorporating multiple predictors lead to an improved predictive accuracy of the SPT. The GB algorithm leads to an accuracy improvement from 0.82 to 0.84 for inhalation allergies, 0.57 to 0.75 for ingestion allergies and 0.69 to 0.80 for all allergies, when compared to the cutoff based model currently used in clinic.

The two studies can be integrated by the development of a tool that automates the SPT result reading and predicts the patient outcome, utilizing the improved prediction model. By implementing the tool into a smartphone device, general practitioners are able to perform the SPT in secondary care. This will lead to a simple and quick diagnosis, resulting in improvement of the quality of care for the patients and in time saving for allergists and nurses, especially when the tool is directly linked to the electronic health record of the patient. Furthermore the study could lead to more conformity of SPT results throughout the country, enabling inter-institutional comparison of SPT results.

Samenvatting

De huidpriktest (HPT) is vaak het eerste middel dat ingezet wordt voor de diagnose van een vermoedelijke allergie. De inter-waarnemer variabiliteit en meetfouten van deze test leiden tot een gebrek aan objectiviteit en reproduceerbaarheid. Dit maakt het lastig om HPT uitslagen te vergelijken tussen verschillende ziekenhuizen. Ook worden er op basis van de uitkomst van de huidpriktest relatief veel patiënten onnodig doorgestuurd voor vervolgonderzoek, zoals een provocatietest. Dit leidt tot onnodige extra kosten en ongemak voor de patiënt. Het doel van dit onderzoek is om de HPT te automatiseren en te decentraliseren en om de uitkomst predictie van de patiënt te verbeteren. Op deze manier kan er mogelijk een meer nauwkeurige, kwantitatieve, objectieve en reproduceerbare allergie diagnose gemaakt worden.

In deze thesis wordt een deep learning netwerk toegepast om het aflezen van de HPT te automatiseren en te decentraliseren. Dit netwerk bepaalt het kwaddel oppervlak uit foto's gemaakt van de onderarm van de patiënt. De foto's zijn voorbewerkt en geannoteerd door de auteur. De annotatie is geverifieerd met drie assistenten van de polikliniek, waarna een deep Residual U-net (ResUnet) is getraind op de training data. De resultaten laten een dice similarity coefficient van 0.77 zien, een intersection-over-union van 0.55 en een nauwkeurigheid van 0.91. Vergeleken met andere computer vision methodes in de literatuur, heeft deze methode een gelijke nauwkeurigheid met een betere precisie.

Voor een verbeterde uitkomst voorspelling van de aanwezigheid van een allergie bij een patiënt, wordt de voorspellende nauwkeurigheid van de HPT getracht te verbeteren via twee strategieën. Allereerst worden de meetfouten die gemaakt worden tijdens het aflezen van de HPT verminderd. De kwaddeloppervlaktes worden in Matlab semi-automatisch bepaald vanuit de papieren uitslagen van de HPT. Daarnaast worden meerdere klinische gegevens van de patiënt toegevoegd in een complexer voorspellingsmodel. Vijf verschillende modellen worden geëvalueerd: de methode die op dit moment gebruikt wordt in de kliniek en vier machine learning methodes: Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB) en Logistic Regression (LR). De resultaten laten zien dat zowel een verbeterde kwaddelgrootte bepaling als een complexer predictiemodel met meerdere klinische voorspellers bijdragen aan een verbeterde nauwkeurigheid in het voorspellen van de patiënt uitkomst. Het GB model zorgt voor de grootste verbetering in nauwkeurigheid: van 0.82 naar 0.84 voor inhalatie allergenen, 0.57 naar 0.75 voor ingestie allergenen en 0.69 naar 0.80 voor alle allergenen.

De twee studies kunnen geïntegreerd worden door een tool te ontwikkelen die zowel de HPT automatiseert als de uitkomst van de patiënt voorspelt met het verbeterde voorspellingsmodel. Door deze tool in een mobiele applicatie te implementeren, kunnen ook huisartsen de HPT uitvoeren in de tweedelijnszorg. Dit leidt tot een simpele en snelle diagnose en daarmee een verbetering van de zorgkwaliteit voor de patiënt. Ook zal het tijd besparen en gemak opleveren voor de verpleegkundigen die nu de test uitvoeren, met name wanneer de tool direct gekoppeld wordt aan het elektronisch patiënten dossier. Tot slot zal het leiden tot meer conformiteit van HPT resultaten, wat landelijke studies en vergelijking van HPT resultaten mogelijk maakt.

Contents

1 Introduction

Atopic diseases, among which food allergies, have strongly increased over the past 50-60 years in many Western countries. The point prevalence of food allergies in Europe is around 6%, indicating that 1 out of 20 children is suffering from a food allergy [1].

The Skin Prick Test (SPT) is the first step in allergy diagnostics and it is a widely used tool [3]. The SPT detects sensitizations to specific allergens. The process of sensitization is explained in Appendix A. The SPT is simple, easy to carry out and the results are available immediately. Besides, by making the results visible, the sensitization of a certain allergen is easy to understand for a child. SPT's are applied to the forearm of the patient. Multiple allergens are introduced simultaneously by placing drops of allergen extracts on the skin. Besides, a negative control extract and a positive histamine control extract are added. Subsequently, the skin underneath the drops is pierced with a small metallic sterile lancet [4]. The procedure is shown in Figure 1. Patients with a hypersensitivity to an antigen will provoke a raised itchy area on the skin at the location of the drop. This area is called a wheal and it is surrounded by erythema. The wheal size of the antigen in relation to the positive control wheal is called the Histamine Equivalent Wheal Size (HEWS) and it indicates the degree of sensitization of the antigen.

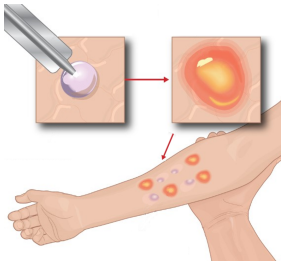


Figure 1. The Skin Prick Test, adjusted image from [2]

To determine the HEWS, the wheals on the skin are marked with a pen. An adhesive tape is used to transfer the markings to a white paper sheet. The size of the wheal is measured in terms of a so-called mean diameter: the longest diameter of the wheal and its perpendicular diameter are summed and divided by two [4]. A positive SPT result includes a HEWS > 0.4 and/or a mean wheal diameter > 3 mm [5, 6]. After a positive SPT, follow-up diagnosis may be necessary to determine whether a patient has a clinical allergy. The total workflow of allergy diagnosis is shown in Figure 2.

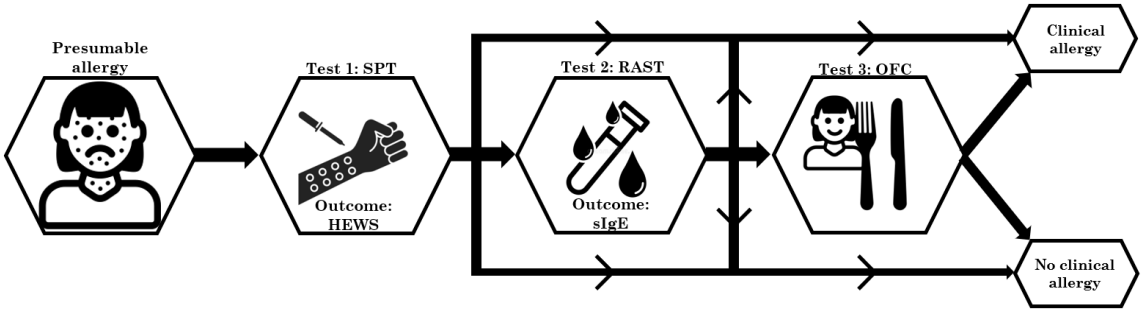


Figure 2. The allergy diagnosis in primary care

The SPT has several limitations. First of all, the drawing of the wheal contour is observer dependable and has a poor reproducibility. Secondly, the time between introducing the allergen and reading the result is variable and not all the wheals can be measured simultaneously. Since wheals wane over time, this introduces an additional variability. On top of these inter-observer variabilities,

measurement errors occur from ink spreading and the fact that that wheals often do not present as circles, leading to an inaccurate mean diameter. The above described lack of objectivity and reproducibility in SPT results make it hard to decentralize the test and compare test results between different institutions. Another limitation of the SPT is that sensitization does not necessarily lead to a clinical allergy, resulting in a relatively high amount of unnecessary follow-up diagnosis and additional costs [7–9].

1.1 Study objective

The ultimate goal is to improve the diagnostic value of the SPT, by decentralizing the test, enabling inter-institutional comparison of SPT results and improving the patient outcome prediction.

In order to fulfill the ultimate goal, the objectives of this thesis are to automate the SPT result reading and to improve the patient outcome prediction. These two components could be combined in a tool that can be used in clinic, potentially leading to a more accurate, quantitative and reproducible diagnosis. The research questions comprising the two objectives are as follows:

- *To what extent can wheal surface areas be (semi)automatically computed from a single picture of the forearm of the patient, using a deep learning model?*
- *Which parameters have a diagnostic predictive value for the outcome of the patient and how can they be implemented in a prediction model?*

1.2 Outline

The research questions are addressed in two parts, provided in paper format. The first paper will focus on the development of the (semi) automatic wheal segmentation algorithm. The second paper contains the comparison of different diagnostic prediction models. Both the papers are preceded with technical background information. The thesis will finalize with an overall discussion and conclusion, in which the clinical impact of the papers is addressed and recommendations for future work are made.

2 Background 1: Deep Residual U-Net

A deep Residual U-net (ResUnet) is a segmentation network that combines the strengths of deep residual learning [10] and the U-net [11]. A U-net is a Convolutional Neural Network (CNN) developed for biomedical image segmentation by Ronneberg et al. The name U-net refers to the shape of the CNN, as shown in Figure 3. The core building block with the most computational contribution in the U-net is the convolutional layer. A convolutional layer performs a linear operation that involves multiplication of the input with a set of weights, organized in a two-dimensional filter. The composition of the filter is initially random, but the weights are optimized during the training of the network. Repeated application of the same filter to the input results in a map of activations, which is called a feature map. The feature map indicates the location and strength of detected features in the input. The feature map is passed through a Rectified Linear Unit (ReLU) activation layer. This layer transforms its input to zero or positive [12].

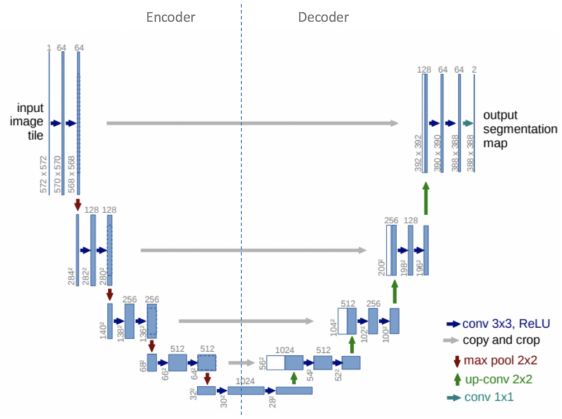


Figure 3. U-net architecture [11]

The U-shape of the U-net derives from the encoder and decoder part of the network. The encoder part includes convolutional blocks followed by a max pooling layer to encode the image into feature representations for the following layer. The encoding path reduces the spatial information and increases the feature information. The decoder part projects the features learnt by the encoder onto the pixel space. The decoder consists of upsampling, concatenation and convolutional operations and it combines the feature and spatial information with skip connections. The advantages of these skip connections include avoidance of the vanishing gradient effect and combination of global information with local information. The vanishing gradient effect occurs when the network is unable to propagate useful gradient information from the output end of the model back to the layers near the input end of the model. This effect is caused by the high amount of layers in a deep neural network [13].

Even though the vanishing gradient problem is solved within the U-net by skip connections, the performance of the deep network often still can get stagnated, due to the degradation problem. The degradation problem means that with the increase of the network layers, the accuracy drops. In other words: if a network is performing best with 10 layers, adding 20 more layers will decrease the performance. The additional 20 layers will have to propagate the same result as the 10th layer by outputting the identity function $f(x) = x$, which is a hard function to learn for a neural network. To overcome this problem, residual learning is introduced. With residual learning, the information from early layers is passed directly to the deep layers, like an identity function. The deep layers

only have to learn the function $f(x) = 0$, which is an easier function to learn than $f(x) = x$ [10]. The differences between the plain building blocks in the U-net and residual units of the ResUnet are shown in Figure 4. Combining the strenghts of residual learning and the U-net leads to two benefits: residual learning eases the training of deep networks and the skip connections of the U-net facilitates information propagation, allowing better performance and avoidance of the vanishing gradient problem [14].

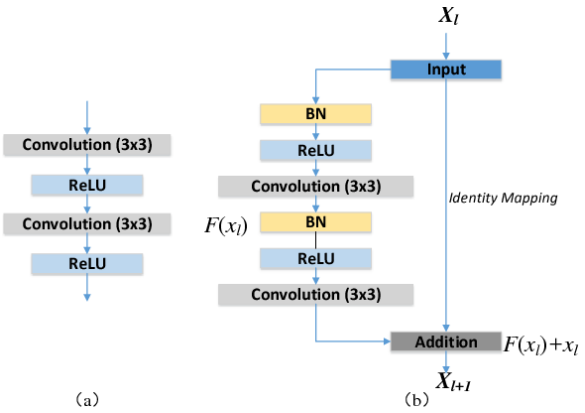


Figure 4. Building blocks of (a) U-Net and (b) ResUnet with identity mapping [14]

3 Paper 1

To be submitted to *Allergy* as an original article

Wheal segmentation by a deep residual U-net: towards (semi-)automatic skin prick test reading

M.S.M. Geessinck[1,2], R.F.M. van Doremalen[1,2], J.E. De Jong[1,2], D.M.W. Gorissen[2], F. van der Heijden[1], J. Faber[2]
[1]Robotics and Mechatronics (RaM), University of Twente, Enschede, Netherlands
[2]Deventer Ziekenhuis, Deventer, Netherlands
E-mail: mariekegeessinck@hotmail.com

Abstract: *Background:* Skin Prick Test (SPT) results contain uncertainties, introduced by the manual way of drawing the wheal contour and determing the wheal size. It is hypothesized that an automated tool potentially reduces these uncertainties and that it may enable decentralization of the SPT from primary to secondary care.
Methods: The proposed method operates with a deep Residual U-net (ResUnet) on photographs from the forearm of patients. The pre-processing pipeline contains color correction, camera perspective correction and contrast enhancement. Data annotation is done by a researcher (MSMG) and verified by three assistants of the outpatients' clinic. The model parameters are optimized with random search and 5 fold cross validation is performed to extract the best performing model.
Results: The best performing model shows a dice similarity coefficient of 0.77, an intersection-over-union of 0.55 and an accuracy of 0.91 on the test set.
Conclusions: The paper proposes a novel deep learning based approach to reduce the measurement errors made in the manual laborious reading process currently used in clinic. The results show a similar accuracy with a higher precision compared to other computer vision based approaches in literature.
Index terms— Allergy test, digital photography, residual U-net, skin prick test, wheal detection

1 Introduction

The skin prick test (SPT) is one of the most commonly used methods for diagnosing allergies for over many decades [1]. During the procedure of the SPT, wheals are marked with a pen and the contours are transferred with translucent adhesive tapes to a white paper sheet [2]. The mean diameter of the wheal is determined in a manual laborious manner, which introduces uncertainties in the test outcomes. The wheal marking is observer dependable and the area determination is inaccurate. Furthermore, some allergists visually inspect the wheals, instead of measure them, which makes the results dependent of qualitative judgement [3–5]. An automated tool would potentially reduce the uncertainties introduced by the reading process and minimize the inter-observer variability. Besides, by standardizing the SPT results, it could be possible to accurately follow the sensibilization of a patient for a certain allergy over time. This may be of interest, since patients can overgrow certain allergens when they become older. Furthermore, an automated SPT reading tool allows comparison of the test between care facilities and could enable transition from primary to secondary care.

This paper is aimed at the development of a (semi-)automatic wheal segmentation method. The proposed method operates by a deep learning network, segmenting the wheal areas from digital photographs taken from the patient's forearm. Even though deep learning based segmentation models for clinical applications have evolved dramatically in the past decade, the possibilities of wheal segmentation using a deep learning model is an unexplored field of study, which makes the proposed method a novel approach.

2 Methodology

2.1 Study design

This prospective single-institution study took place at Deventer Ziekenhuis, which is specialized in allergy treatment. The study population is aged under 18 years old. All participants are scheduled for a SPT and the patients and gave verbally consent for taking a photo of the forearm during the test. All images are processed anonymously. The inclusion and exclusion criteria for this study are mentioned below.

Inclusion criteria:

- Children (age $\geq$ 1 and $\leq$ 18 years) who are referred for a SPT
- SPT performed on forearm
- Able to hold forearm in position on table for 3 seconds

Exclusion criteria:

- No consent
- Anxious or nervous child
- Failure of the SPT (positive control does not induce a wheal or negative control induces a wheal)

2.2. Data acquisition and pre-processing

The data is collected with an Apple iPad Pro 2020, with use of an Aruco marker and a MacBeth ColorChecker. The iPad is kept parallel to the arm of the patient and the distance between the iPad and the arm is 25 cm. The protocol of the data collection is attached in Appendix B. In order to standardize the

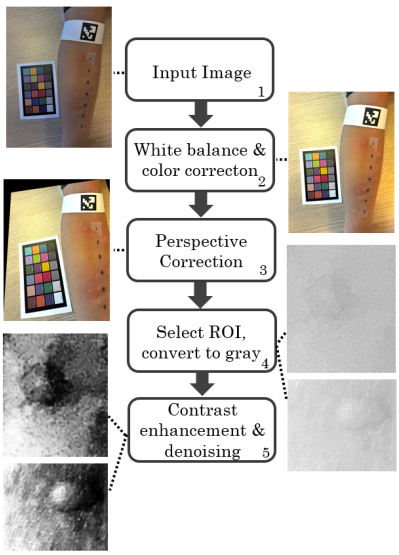


Fig. 1. The pre-processing pipeline

SPT, some pre-processing steps are required to correct for different illumination conditions. The collected photo is corrected for white balance and color correction with the ColorChecker is performed. Afterwards, the camera perspective is corrected, with use of an Aruco marker. After perspective correction, a Region Of Interest (ROI) around the wheal is selected and converted to grayscale. The contrast in the image is enhanced and the image is denoised, respectively with histogram equalization and Gaussian filtering (kernel standard deviation: 0.9). The pre-processing pipeline is shown in Fig. 1. Information about the color correction procedure is attached in Appendix C.

2.3. Data annotation

The data is annotated by a researcher (MSMG), after training by a specialist (DMWG). A subset of 20% of the annotations is verified by three assistants of the outpatients' clinic in a not fully crossed design [6]. Three annotator pairs are formed, each consisting of the researcher and one of the assistants of the outpatients' clinic. The Inter-Annotator Reliability (IAR) is measured with the Light's kappa, which allows for three or more raters and corrects for chance agreement. Table 1 gives an interpretation of the kappa and the formula's for calculation of the kappa are provided in Appendix D.

Value of Kappa	Level of Agreement	% of data that are Reliable
0 - .20	None	0 - 4%
.21 - .39	Minimal	4 - 15%
.40 - .59	Weak	15 - 35%
.60 - .79	Moderate	36 - 63%
.80 - .90	Strong	64 - 81%
Above .90	Almost Perfect	82 - 100%

Table 1. Interpretation of kappa [7]

2.4. Data subsets

20% of the total data is used for testing. The residual samples are divided into 80% training data and 20% validation data. The training set will be augmented to 3000 samples by applying a random combination of the following five modifications: (1) width shift range, fraction of 0.3 of total image width, (2) height shift range, fraction of 0.3 of total image width, (3) zoom range of 0.8 to 1.2 of the image size, (4) horizontal flip, (5) vertical flip.

2.5. Deep Residual U-net

This paper uses a deep Residual U-net (ResUnet) as proposed by Zhang et al., which is a segmentation network that combines the strengths of deep residual learning and the U-net [8–10]. The network follows the architecture of the U-net, but uses residual units instead of plain neural units as building blocks. These residual units will ease the training of the network. The skip connections within these residual units and between the low and high levels of the network help avoidance of the vanishing gradient problem and design a network with much fewer parameters. The network architecture is shown in Fig. 2. Some adjustments are made to the proposed model: instead of using the loss function Mean Squared Error (MSE), this paper uses the Dice Loss as loss function. The Dice

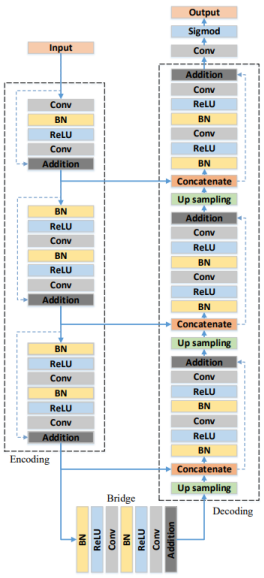


Fig. 2. The proposed ResUnet architecture [8]

loss is commonly used in the medical domain, because it can handle with class imbalanced datasets. Furthermore, Dropout layers are added after each convolutional layer, which reduces overfitting and improves the generalization of the model.

2.6. Hyperparameter optimization
Hyperparameter optimization of the deep ResUnet is performed with the random search method. For each hyperparameter a range of values is defined. These ranges together form a grid in which randomly a combination is chosen to train the model with. Hyperparameters that are optimized are the learning rate, the amount of filters and dropout rate.

2.7. Model evaluation
After hyperparameter optimization, 5 fold cross validation is performed on the training set to extract the best performing model. For each fold, three executions are performed. The best model of the three executions is saved. The saved models of the 5 folds are compared and the best model is selected and tested on the test set. The model performances are reported in the Intersection-Over-Union (IOU) Dice Similarity Coefficient (DSC) and accuracy, which are calculated as follows:

$$IOU = \frac{TP}{TP + FP + FN} \tag{1}$$

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{2}$$

$$Acc = \frac{TN + TP}{FN + FP + TN + TP} \tag{3}$$

In which FN is the amount of false negative pixels, FP is the amount of false positive pixels, TN is the amount of true negative pixels and TP is the amount of true positive pixels.

3 Results

3.2. Data annotation
74 images were collected and annotated. The calculated Light's Kappa is 0.86. According to Table 1, this means that there is a strong level of agreement and that 74% of the data is reliable. The results for each annotator pair is shown in Table 2.

Annotator pair	κ
1	0.87
2	0.89
3	0.84

Table 2. Cohen's kappa per annotator pair

3.3. Hyperparameter Optimization
The results of hyperparameter optimization with grid search are shown in Table 3.

Hyperparameter	Range	Best value
Learning rate	$[1e^{-3}, 3e^{-3}, 5e^{-3}, 8e^{-3}, 1e^{-2}, 1.5e^{-2}]$	$3e^{-3}$
Dropout rate	[0.3, 0.5, 0.8]	0.5
Filters	[[8, 16, 32, 64, 128], [16, 32, 64, 128, 256], [32, 64, 128, 256, 512]]	[8, 16, 32, 64, 128]

Table 3. Hyperparameters that are optimized and their best performing value

3.4. Model performance

The results of 5 fold cross validation are shown in Table 4.

	DSC	**IOU**	**Acc**
Fold 1	0.76	0.54	0.93
Fold 2	0.82	0.67	0.93
Fold 3	0.78	0.68	0.89
Fold 4	0.82	0.63	0.94
Fold 5	0.70	0.49	0.86
Average	0.78 ±0.05	0.60 ±0.08	0.91 ±0.03

Table 4. Scores of 5 fold cross validation

The best performing model from fold 2 is extracted and tested on the test set. Some of the results are shown in Fig. 3 and the residual test images and their results are attached in Appendix E. The DSC on the test set is **0.77**, the IOU is **0.55** and the accuracy is **0.91**.

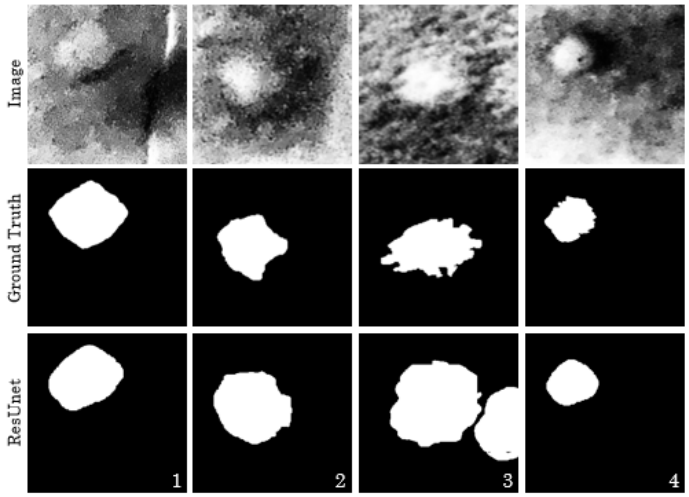


Fig. 3. Results on 4 test samples: image (top), ground truth (middle), result ResuNet (bottom)

4 Discussion

For interpretation of the obtained results, the performances are compared to related work in literature. Two studies were found with the aim of wheal segmentation. Bulan et al [1] show an average accuracy of 94% and another paper of Bulan et al. [11] presents an average accuracy of 90%, with

a standard deviation of 30%. Unfortunately, DSC / IOU scores are not provided in those papers, which makes comparison challenging, since the segmentation accuracy highly depends on the percentage of wheal area within the ROI. However, based on these numbers, it can be stated that this paper proposes a method with a comparable accuracy and higher precision. To gain insight in the minimally required performances of the model, a definition of clinically acceptable segmentation performance is required, which has yet to be defined. Several methods for measuring the wheal size are described in literature [1–3, 11, 12]. Each of these methods possesses its own limitations, but the main limitations comprise that they are hard to decentralize, labor-intensive and require expensive equipment. Taken the performances and practicability together, it can be stated that the proposed method outperforms related work. The proposed method can contribute to a user friendly diagnostic tool, reducing the errors introduced by the manual way of performing the SPT. It will ease the performance of the SPT for the nurses and will save time, since the method does not require manually wheal drawing and measurement. Furthermore, it might enable decentralization of the SPT from primary to secondary care and it allows monitoring the sensibilization of patients over years, to trace a possible tolerance development.

To further increase the network's performance and its ability to generalize, it is of interest to enlarge the training data, since a large training set with a wide variety helps to learn common features among different subjects. For example, big wheals appear white, whereas very small wheals appear dark (see appendix E or Fig. 1 for the different color appearance of the wheals). The network has been presented more white appearing wheals compared to dark appearing wheals within the training. As a result, the network is not performing well on the appearing dark wheals. Furthermore, only white patients have been included in this study. Therefore, a recommendation for future work is to investigate the performance of the segmentation algorithm on patients with dark skin color.

5 Conclusion

This paper presents a novel deep learning based approach to read SPT results. A Residual U-net is successfully trained and presents a similar accuracy with a higher precision, compared to other computer vision based algorithms. Taken the performances and practicability together, it can be stated that the proposed method outperforms related work.

References

[1] O. Bulan, "Improved wheal detection from skin prick test images," in *Image Processing: Machine Vision Applications VII*, vol. 9024. International Society for Optics and Photonics, 2014, p. 90240J.

[2] S. Wöhrl, K. Vigl, M. Binder, G. Stingl, and M. Prinz, "Automated measurement of skin prick tests: an advance towards exact calculation of wheal size," *Experimental dermatology*, vol. 15, no. 2, pp. 119–124, 2006.

[3] X. Justo, I. Díaz, J. Gil, and G. Gastaminza, "Prick test: evolution towards automated reading," *Allergy*, vol. 71, no. 8, pp. 1095–1102, 2016.

[4] R. Vieira dos Santos, R. G. Titus, and H. Cavalcante Lima, "Objective evaluation of skin prick test reactions using digital photography," *Skin Research and Technology*, vol. 13, no. 2, pp. 148–153, 2007.

[5] "IgE in Clinical Allergy and Allergy Diagnosis — World Allergy Organization." [Online]. Available: https://www.worldallergy.org/education-and-programs/education/allergic-disease-resource-center/professionals/ige-in-clinical-allergy-and-allergy-diagnosis

[6] K. A. Hallgren, "Computing inter-rater reliability for observational data: an overview and tutorial," *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.

[7] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[8] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[11] O. Bulan and Y. Artan, "Wheal detection from skin prick test images using normalized-cuts and region selection," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1250–1253.

[12] C. Svelto, M. Matteucci, A. Pniov, and L. Pedotti, "Skin prick test digital imaging system with manual, semiautomatic, and automatic wheal edge detection and area measurement," *Multimedia Tools and Applications*, vol. 77, no. 8, pp. 9779–9797, 2018.

4 Background 2: Machine learning classifiers

Machine learning models can be used to give a probability of the presence or absence of a certain allergy. Choosing the most optimal performing model for this classification task, depends on the data. This study composes continuous numerical data (wheal size, wheal circumference, HEWS, HEWC and sIgE), discrete numerical data (age) and categorical data (type of allergy). Combination of these data characteristics and the aim of the study led to the following model requirements:

- Suitable for categorical and numerical data
- Able to deal with missing values
- Interpretable (avoidance of black box models)

Delgado et al. evaluated 179 classifiers based on 17 classifier groups [15]. Their top 10 models derive from Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Boosting (BST) classifier groups. SVM classifiers and NN's do not meet the requirement of interpretability and are therefore discarded. The other two groups will be explained in the sections below. Furthermore, it is chosen to include the Logistic Regression (LR) classifier, since this classifier is easy to implement, interpret and efficient to train [16].

4.1 Random Forest

Machine learning models can be fitted to data individually, or combined in an ensemble. An ensemble is a combination of simple individual models that together create a more powerful model. RF is such an ensemble in which weak learners (decision trees) are combined to form a strong learner with more stability and a higher accuracy [17]. The name RF refers to the randomness which is added twice in the model. First, random subsets of all the data are divided over the decision trees. Subsequently, instead of searching for the most important feature while splitting a node within these decision trees, RF's choose a random subset of the variables at each node and finds the best variable and value from this subset [18]. The different trees perform independently in parallel and the final classification is based on majority voting of the outputs of the trees. The working principle of RF's is shown in Figure 5.

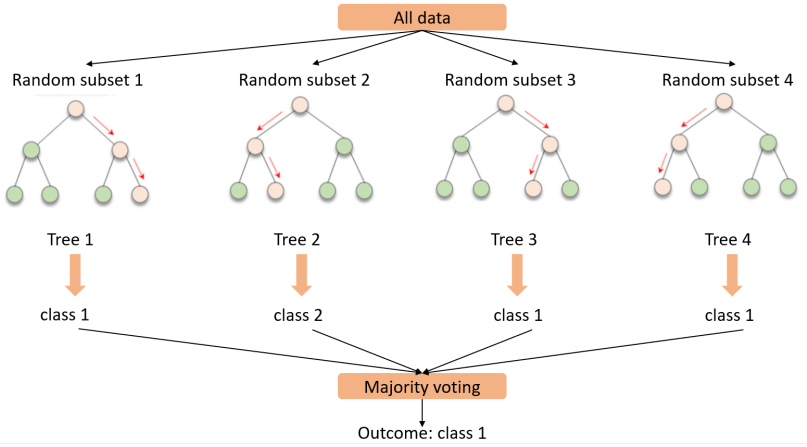


Figure 5. Random Forest working principle, adjusted image from [19]

Within a tree of the Random Forest, each node contains a Gini impurity. The Gini impurity is the probability that a randomly chosen sample in a node would be incorrectly labeled if it was labeled by the distribution of samples in the node. For example: if a node contains 6 samples, of which 2 belong to class 0 and 4 belong to class 1, the Gini impurity is calculated as follows:

$$I_G(n) = 1 - \sum_{i=1}^{J}(p_i)^2 = 1 - ((\frac{2}{6})^2 + (\frac{4}{6})^2) = 0.44 \quad (1)$$

The Gini impurity decreases further down the tree. The most relevant features result in a large drop of the Gini impurity between two nodes and less relevant features result in less decrease of the Gini impurity. This principle can be used for feature selection. In RF feature selection, a feature importance is determined for each feature. This feature importance is the average impurity decrease caused by a question asked for the specific feature. The features with the highest feature importances are used to train the model and other features could be discarded [20].

4.2 Extra Trees

Extra Trees (ET) is another ensemble method, which is in various ways similar to the RF and also combines multiple decision trees to a stronger predictor. The main differences are the following: RF's subsample the input of each tree with replacement, whereas ET use the whole original data set and select subsets without replacement for the input for each tree. Another difference is the selection of cut points for the split of each nodes. Whereas in RF the local optimal feature-split combination is chosen, the ET algorithm selects a random value for the split for a feature being considered, which leads to more diversified trees. An overview of the main differences between decision trees, RF and ET is shown in Table 1 [21, 22].

	Decision Tree	Random Forest	Extra Trees
Number of trees	1	Many	Many
Number of features considered for split at each decision node	All features	Random subset of features	Random subset of features
Sample drawing	Not applied	With replacement	Without replacement
How split is made	Best split	Best split	Random split

Table 1. Differences between Decision Tree, Random Forest and Extra Trees

4.3 Gradient Boosting

Gradient Boosting (GB) is also a tree based ensemble. It combines weak learners sequentially, so that each new tree corrects the error of the previous one. The performance of first tree is expressed in a loss function and boosting relies on the intuition that the best possible next model, combined with the previous models, minimizes the overall loss. The idea of GB relies on filtering observations, leaving those observations that the weak learner can handle and focusing on developing new weak learners to handle the remaining difficult observations. Each next tree is refocused on the examples that the previous ones found difficult and misclassified [23, 24]. An example of the working principle is shown in Figure 6.

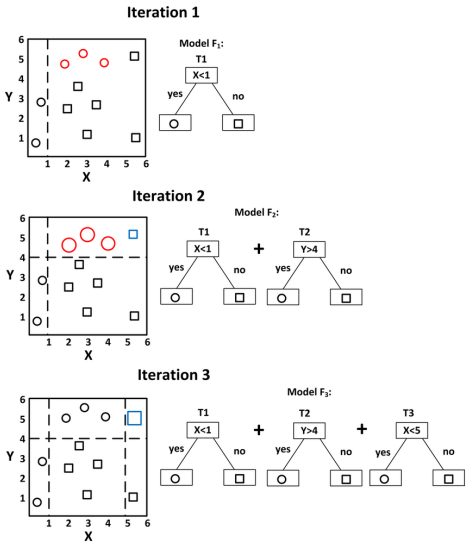


Figure 6. Gradient Boosting working principle [25]

4.4 Logistic Regression

Logistic Regression (LR) predicts the probability of a certain condition being true or false. It fits a S-shaped curve of the probability, as shown in Figure 7.

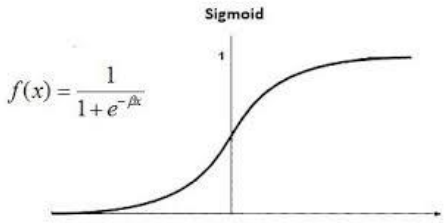


Figure 7. Logistic Regression working principle [26]

Often it is stated that if the probability is $> 50\%$, the condition is true and otherwise, it is false. Figure 7, shows the regression formula with only one feature. In the case of multiple features, the regression formula is as follows:

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} \tag{2}$$

Ordinary least squares regression finds the coefficients of the above described formula by minimizing the squared prediction errors across the training data, represented by:

$$\sum_{i}^{n} (y_i - \widehat{y}_i)^2 \tag{3}$$

The more variables that are included in the regression, the more likely it is to run into excessive covariance. To reduce the chance of overfitting, the variance of the model can be reduced. This

is done with regularization. There are two types of regularization: Lasso and Ridge. Both the regularization methods work by adding a new term to the cost function that is used to derive the regression formula.

Lasso adds the sum of the magnitudes of all the coefficients in the model:

$$\sum_{i}^{n} (y_i - \widehat{y}_i)^2 + \lambda \sum_{j}^{p} \|\beta_j\| \tag{4}$$

Ridge follows the same pattern, but the penalty term is the sum of the coefficients squared:

$$\sum_{i}^{n} (y_i - \widehat{y}_i)^2 + \lambda \sum_{j}^{p} \left\|\beta_j^2\right\| \tag{5}$$

Including the extra penalty term obstructs including extra features. An extra feature may help the regressor minimize the first term of the cost function, but it will increase the penalty term. This way, there is a balance in which the value of the increasing coefficient is weighted against the corresponding increase to the overall variance of the model, resulting in only using the most valuable features. Ridge and Lasso act as their own sort of feature selection: the coefficients of the features which don't have much predictive powered are pushed down, while the more predictive features obtain higher coefficients. Ridge regression will never push a coefficient all the way down to zero, since it is squares the coefficients. Lasso is able to make some of the coefficients zero [27, 28]. The presence of Lasso and/or Ridge regularizations and the value of λ are optimized during Grid Search.

5 Paper 2

To be submitted to *Allergy* as an original article

Machine learning models for allergy prediction: improving the clinical value of Skin Prick Test results

M.S.M. Geessinck[1,2], R.F.M. van Doremalen[1,2], D.M.W. Gorissen[2], F. van der Heijden[1], J. Faber[2]
[1]Robotics and Mechatronics (RaM), University of Twente, Enschede, Netherlands
[2]Deventer Ziekenhuis, Deventer, Netherlands
E-mail: mariekegeessinck@hotmail.com

Abstract: *Background:* Uncertainties in the Skin Prick Test (SPT) results, together with an increasing prevalence of allergies and an increasing number of patients seeking for diagnosis, lead to a high amount of unnecessary oral food challenges (OFCs). This raises the need for an improved accuracy in predicting the OFC outcome. It is hypothesized that a more accurate wheal size determination and an improved predictive model could enlarge the diagnostic value of the skin prick test by improving its predictive accuracy. This way, the amount of unnecessary OFCs could potentially be reduced, leading to less costs and discomfort for patients and their parents.
Methodology: 305 SPT results were digitized and the wheal area, wheal circumference, area/circumference ratio, histamine equivalent wheal size (HEWS) and histamine equivalent wheal circumference (HEWC) were extracted. Patient specific information among which age, sIgE levels, presence of asthma, presence of rhinitis and presence of eczema were extracted from the electronic health record. The cutoff based model as currently used in clinic is compared to four machine learning models: Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB) and Logistic Regression (LR).
Results: The GB classifier shows an improvement of predictive accuracy from 0.82 to 0.84 for inhalation allergies, 0.57 to 0.75 for ingestion allergies and 0.69 to 0.80 for all allergies.
Conclusions: This study shows that a more complex classifier, with additional predictors and a more accurate wheal size determination, could improve the clinical value of skin SPT results by improving its predictive accuracy.
Index terms— Skin prick test, Machine learning, allergy diagnosis, classification models

1 Introduction

Food allergies have shown both an increasing prevalence and an increasing number of patients and parents seeking diagnosis [1–4]. The gold standard for the diagnosis is the oral food challenge (OFC). However, this procedure is time-consuming, costly and patients and their parents may have fear of risk of sever systemic reactions during the OFC [5]. Given these factors, diagnostic tests have been developed to predict the OFC outcome. These tests focus on cutoff values for serum-specific IgE (sIgE) levels and for Skin Prick Test (SPT) results [6]. However, the SPT results are determined in a manual, laborious manner and cutoff values of sIgE are highly influenced by the age of the children, resulting in outcomes with uncertainties. This results in a need for an improved accuracy in predicting the OFC outcome and thereby potentially reducing the number of OFCs.

This paper is aimed at improving the predictive accuracy of the OFC outcome considering two strategies. Initially, the SPT measurement errors are reduced by development of a semi-automatic algorithm that extracts wheal characteristics from digitized SPT results. Secondly, multiple clinical factors are incorporated into a more complex predictive model. The classification method currently used in clinic is compared with four machine learning based classifiers: Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB) and Logistc Regression (LR). The current method in clinic classifies a SPT outcome as positive if the manually determined wheal size of the antigen in relation to the postitive control (HEWS) > 0.4 or mean diameter > 3 mm. Additionally to the comparison of the different classification methods, this study also investigates the potency of incorporating more clinical predictors in the predictive model, compared to only using the HEWS as predictor. These predictors comprise the wheal area, wheal circumference, wheal circumference of the antigen in relation to the positive control (Histamine Equivalent Wheal Circumference (HEWC)), sIgE values, age, presence of rhinitis, presence of eczema and presence of asthma.

2 Methodology

2.1 Study design
This retrospective single-institution study took place at Deventer Ziekenhuis, which is specialized in allergy treatment. SPT results from the period 2017 - 2018 were analyzed. The study population is aged under 18 years old. All paper sheets were made anonymously before scanning. The inclusion and exclusion criteria are mentioned below.

Inclusion criteria:
- Children (age ≥ 1 and ≤ 18 years)
- SPT performed in Deventer Ziekenhuis in the period 2017 - 2018

Exclusion criteria:
- Failure of the SPT (positive control does not induce a wheal or negative control induces a wheal)
- No patient outcome corresponding to the SPT found in the EHR

2.2. Data pre-processing
A Matlab script is made to segment the wheal and extract the wheal area, wheal circumference, area/circumference ratio, HEWS and HEWC from the digitized paper sheets. The HEWS is calculated and stored in twofold. The first measure is according to the current manually method of

measuring the wheal size in relation to the positive control performed by the allergist, in which the wheal size is approached with a mean diameter. The second measure is determination of the wheal size in relation to the positive control, in which the wheal area is determined in pixels with the Matlab script. An example of the data extraction from a SPT result is shown in Fig. 1 and the code to achieve this result is attached in Appendix F. The age, sIgE level and information about the presence of asthma, presence of rhinitis and presence of eczema are extracted from the Electronic Health Records (EHR) of the patients.

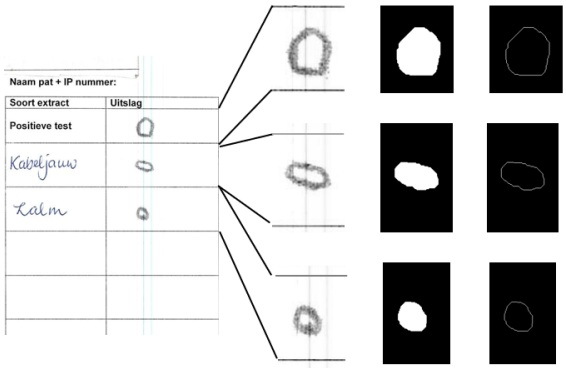


Fig. 1. Extraction of wheal information from digitized SPT results

2.3. Feature selection

First of all, the missing values within the collected data are handled. A missing rate of 10% or higher will result in biased results [7]. After imputation of the missing values, feature selection is performed to remove the irrelevant features. This way, irrelevant features do not longer contribute to the computational cost of the model and do not have a negative influence on the performance of the mode. Due to the good interpretability, high accuracy and good generalizability, it is chosen to use the RF for embedded feature selection [8, 9]. Within the RF, the feature importances are calculated by the average Gini impurity decrease caused by a question asked from each feature [10]. The features responsible for 95% of the total importance will be selected as the inputs of the model.

2.4. Hyperparameter optimization

The total data is split into 75% training set and 25% test set. Within the training set, 4 fold cross validation is performed during parameter optimization with the random search method. The probability of at least one of the best p% sets of parameters drawn from the grid after n iterations is:

$$P = 1 - (1 - p)^n \tag{1}$$

Solving the number of iterations for random search, having the best 1% sets of parameters with 99% confidence:

$$0.99 = 1 - (1 - 0.01)^n, n \approx 460 \tag{2}$$

The hyperparameters that are optimized and their ranges are shown in Table 2 in Appendix G.

2.5. Model evaluation

After hyperparameter optimization, the optimal parameters are selected and the model performances

are evaluated on the test set. The performance of the model are reported in the accuracy and the precision:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

3 Results

3.1. Data

305 SPT's were digitized, resulting in 769 wheals, of which 382 inhalation wheals and 387 ingestion wheals. The patient information is divided into three datasets: All wheals (with type of allergy as an additional parameter), inhalation wheals and ingestion wheals. It is chosen to make this subdivision, since the pathophysiology of inhalation and ingestion allergies differ, which may require a different type of predictive model.

3.2. Feature selection

59% of the sIgE values were missing, since not every patient with a SPT undergoes the radioallergosorbent test. Within the inhalation and ingestion datasets, this was resp. 69.4% and 48.8%. These ratios are too high for a reliable imputation and therefore the sIgE parameter is discarded. The feature importances for each variable is shown in Table 1. The table shows that all parameters excluding rhinitis, eczema and asthma are responsible for 95% of the total importance. This indicates that the models may profit from removing those variables.

Variable	All wheals	Inhalation wheals	Ingestion wheals
HEWS	0.18	0.26	0.18
HEWC	0.16	0.23	0.18
Area	0.14	0.10	0.17
Area/Circumference	0.14	0.24	0.17
Circumference	0.12	0.12	0.16
Type	0.10	-	-
Age	0.10	0.09	0.10
Rhinitis	0.02	0.03	0.01
Asthma	0.01	0.01	0.01
Eczema	0.01	0.01	0.01

Table 1. Feature importances from RF

Based on Table 1, different combinations of features are made for each of the three datasets and these feature groups are shown in Table 2. Feature group 1 contains the features responsible for 95% of the feature importances. Feature group 2 contains the top three most important features. Feature group 3 contains the HEWS calculated with the Matlab algorithm. Feature group 4 contains the HEWS determined manually by the allergist.

Patient group	**Feature group 1**	**Feature group 2**	**Feature group 3**	**Feature group 4**
All allergies	HEWS, HEWC, area, circumference, area/circumference, age, type of allergy	HEWS, HEWC, area, area/circumference	HEWS	HEWS determined by allergist
Inhalation	HEWS, HEWC, area, circumference, area/circumference, age	HEWS, HEWC, area/circumference	HEWS	HEWS determined by allergist
Ingestion	HEWS, HEWC, area, circumference, area/circumference, age	HEWS, HEWC, area, area/circumference	HEWS	HEWS determined by allergist

Table 2. Overview of divided feature groups per patient group

3.3. Hyperparameter optimization

The optimal parameters for each data set and feature group after Random Search per model are shown in Appendix G. The obtained accuracy is an average accuracy of the 4 fold cross validation within the training set. The deviation of the folds is also shown. The optimal parameters are used for the models to test the model performances on the test set.

3.4. Model performances

First of all, the predictive accuracy of the OFC outcome, as performed currently in the clinic is determined. Within this experiment, the OFC outcome was considered as positive if the HEWS or mean diameter as determined manually by the allergist is higher than the aforementioned cutoff values. The predictions were compared with the true OFC outcomes as reported in the EHR of the patient. The results are shown in Table 3

Patient group	**Acc**	**Prec**
All allergies	0.69	0.66
Inhalation	0.82	0.82
Ingestion	0.57	0.51

Table 3. Accuracy and precision of the classification method used currently in clinic

Secondly, the predictive accuracy of the OFC outcome, predicted with the four machine learning based classifiers is determined. For each patient group and each subset of features, the accuracy and precision are determined for all the four models. The results are shown in Table 4. To select the best feature group for each patient group, the mean of the models is calculated and shown in the last column. The results per feature group for each patient group is shown in Fig. 2.

		RF		**ET**		**GB**		**LR**		**Mean of models**	
Patient group	**Model features**	**Acc**	**Prec**	**Acc**	**Prec**	**Acc**	**Prec**	**Acc**	**Prec**	**Acc**	**Prec**
All allergies	Feature group 1	0.80	0.81	0.80	0.82	0.80	0.82	0.79	0.77	**0.80**	**0.81**
	Feature group 2	0.75	0.74	0.74	0.72	0.75	0.77	0.73	0.74	0.74	0.74
	Feature group 3	0.75	0.76	0.76	0.74	0.75	0.75	0.78	0.77	0.76	0.76
	Feature group 4	0.73	0.72	0.73	0.72	0.73	0.72	0.72	0.77	0.73	0.73
Inhalation	Feature group 1	0.84	0.88	0.83	0.84	0.84	0.85	0.82	0.85	**0.83**	**0.86**
	Feature group 2	0.83	0.85	0.84	0.86	0.78	0.81	0.84	0.87	0.82	0.85
	Feature group 3	0.80	0.82	0.80	0.86	0.76	0.81	0.81	0.90	0.79	0.85
	Feature group 4	0.79	0.82	0.78	0.80	0.79	0.85	0.79	0.82	0.79	0.82
Ingestion	Feature group 1	0.71	0.69	0.70	0.69	0.72	0.71	0.73	0.76	0.72	0.71
	Feature group 2	0.72	0.71	0.70	0.68	0.75	0.74	0.74	0.77	**0.73**	**0.73**
	Feature group 3	0.71	0.68	0.73	0.78	0.71	0.68	0.70	0.77	0.71	0.76
	Feature group 4	0.67	0.69	0.72	0.76	0.69	0.72	0.72	0.89	0.70	0.77

Table 4. Accuracy and Precision of the Random Forest, Extra Trees, Gradient Boosting and Logistic Regression, for each patient group and each subset of features

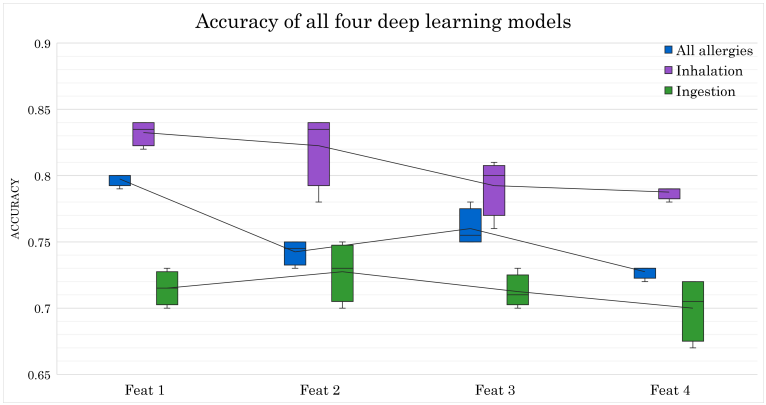


Fig. 2. Accuracy of the four deep learning model per feature group

To simplify Table 4, the best features groups are extracted, based on the 'Mean of models' column and Fig. 2. This is shown in Table 5.

		RF		ET		GB		LR	
Patient group	**Best feature group**	**Acc**	**Prec**	**Acc**	**Prec**	**Acc**	**Prec**	**Acc**	**Prec**
All allergies	Feature group 1	0.80	0.81	0.80	0.82	0.80	0.82	0.79	0.77
Inhalation	Feature group 1	0.84	0.88	0.83	0.84	0.84	0.85	0.82	0.85
Ingestion	Feature group 2	0.72	0.71	0.70	0.68	0.75	0.74	0.74	0.77
Average		0.79	0.80	0.78	0.78	**0.80**	**0.80**	0.78	0.80

Table 5. Accuracy and precision for the four models, given the best model features extracted from Table 4 and Fig. 2

To investigate whether a more accurate wheal size determination results in a better predictive accuracy, a comparison between feature group 3 and 4 is made. The results of these feature groups are extracted from Table 4 and shown in Table 6.

Patient group	**Feature group**	**RF**	**ET**	**GB**	**LR**
All allergies	Feature group 3	**0.75**	**0.76**	**0.75**	**0.78**
	Feature group 4	0.73	0.73	0.73	0.72
Inhalation	Feature group 3	**0.80**	**0.80**	0.76	**0.81**
	Feature group 4	0.79	0.78	**0.79**	0.79
Ingestion	Feature group 3	**0.71**	**0.73**	**0.71**	0.70
	Feature group 4	0.67	0.72	0.69	**0.72**

Table 6. Comparison of the accuracies from feature group 3 and 4 for the four models

The T-test is performed between feature group 3 and 4 for each patient group. The T-values are the following: 4.3 for all allergies, 0.44 for inhalation allergies and 0.91 for ingestion allergies. Considering $\alpha = 0.05$, this means that the difference between feature group 3 and feature group 4 is only statistically significant for the patient group 'all allergies'.

4 Discussion

Comparing feature group 4 to the other feature groups in Table 4, it can be stated that predicting the OFC outcome with additional predictors results in an improved accuracy, compared to only

using the HEWS determined by the allergist as a predictor. The feature combination resulting in the best performances of the four prediction models are determined by looking at the mean of the four models in Table 4 and in Fig. 2. It can be seen that feature group 1 results in the best outcome of the prediction models for all alergies and for inhalation allergies. For ingestion allergies, feature group 2 results in the best predictions. Given these feature groups, it can be stated that the GB model has the best performances, according to Table 5.

For comparison of the results from the GB model to the results from the cutoff based model currently used in clinic, the results in Table 3 and Table 5 can be studied. The results show an improvement of predictive accuracy from 0.82 to 0.84 for inhalation allergies, 0.57 to 0.75 for ingestion allergies and 0.69 to 0.80 for all allergies. This improved accuracy could potentially reduce the amount of unnecessary OFCs and thereby reduce the costs and discomfort of an OFC for patients and their parents.

To assess the influence of an improved wheal area determination, the differences can be studied between the performances based on the HEWS determined by the Matlab algorithm and the HEWS manually determined by the allergist, shown in Table 6. Feature group 3 outperforms feature group 4 in 10 out of 12 cases. This means that the wheal size determination with the Matlab algorithm results in a better predictive accuracy. However, this difference is only statistically significant within the all allergies group.

Even though the results seem satisfactory, there are several aspects that need to be considered. First of all, the performances of the feature group 'HEWS manually determined by allergist' are evaluated on a smaller amount of samples. Ideally, the performances are all evaluated over the same amount of samples. However, not every digitized SPT result contained a written HEWS value determined by the allergist. In these cases, the allergist determined the SPT outcome by visually inspection. Secondly, the results show that a large improvement in accuracy and precision is obtained with the GB model compared to the cutoff value model. This means that more OFC outcomes are correctly predicted and the ratio of false positive outcomes is reduced, potentially resulting in less OFCs. However, this comes with the cost of a decrease in the recall. This means that for some patients, the OFC outcome is false negative predicted. This could especially be dangerous for patients with ingestion allergies, since those allergic reactions can result in sever systemic reactions. Furthermore, this study is a one institute retrospective study in which the data are collected over 1 year. In order to improve the reliability of the results, the data set should be enlarged and the conclusions drawn in this study could be validated with SPT results from other institutions, since it is known that the SPT performance differ per institution.

5 Conclusion

This paper compares the performances of four machine learning classification models incorporating additional clinical predictors with the current method used in clinic to predict the OFC outcome. It shows that both an improved wheal size determination and a more complex predictive model incorporating multiple predictors lead to an improved clinical value of the SPT by improving its predictive accuracy.

References

[1] C. R. Simpson, J. Newton, J. Hippisley-Cox, and A. Sheikh, "Incidence and prevalence of multiple allergic disorders recorded in a national primary care database," *Journal of the Royal Society of Medicine*, vol. 101, no. 11, pp. 558–563, 2008.

[2] D. Hughes and C. Mills, "Food allergy: a problem on the increase." *Biologist (London, England)*, vol. 48, no. 5, pp. 201–204, 2001.

[3] R. Gupta, A. Sheikh, D. Strachan, and H. R. Anderson, "Increasing hospital admissions for systemic allergic disorders in england: analysis of national admissions data," *Bmj*, vol. 327, no. 7424, pp. 1142–1143, 2003.

[4] R. Gupta, A. Sheikh, D. P. Strachan, and H. R. Anderson, "Time trends in allergic disorders in the uk," *Thorax*, vol. 62, no. 1, pp. 91–96, 2007.

[5] S. Roberts, "Challenging times for food allergy tests," *Archives of disease in childhood*, vol. 90, no. 6, pp. 564–566, 2005.

[6] A. DunnGalvin, D. Daly, C. Cullinane, E. Stenke, D. Keeton, M. Erlewyn-Lajeunesse, G. C. Roberts, J. Lucas, and J. O. Hourihane, "Highly accurate prediction of food challenge outcome using routinely available clinical data," *Journal of allergy and clinical immunology*, vol. 127, no. 3, pp. 633–639, 2011.

[7] D. A. Bennett, "How can i deal with missing data in my study?" *Australian and New Zealand journal of public health*, vol. 25, no. 5, pp. 464–469, 2001.

[8] U. Stańczyk, "Feature evaluation by filter, wrapper, and embedded approaches," in *Feature Selection for Data and Pattern Recognition*. Springer, 2015, pp. 29–44.

[9] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[10] "Feature Selection Using Random forest — by Akash Dubey — Towards Data Science." [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

6 Discussion

In the previous chapters, short discussions have already been made. This section will answer the research questions as mentioned in the study objective and will address the integration of the two studies and the clinical value they can create. Furthermore, recommendations for future work will be made.

Paper 1 shows that wheal areas can be (semi)automatically segmented from a single picture of the forearm of the patient, using a deep learning model. The wheal areas are extracted in pixel size, but wheal surface areas are not provided in absolute measures. However, since the HEWS is dimensionless, this is sufficient to determine the HEWS of an allergen. The proposed method in paper 1 can contribute to a user friendly diagnostic tool that is easy to carry out, reproducible and works accurate. It could ease the performance of the SPT for the nurses and might allow decentralization of the SPT from primary to secondary care.

Paper 2 shows that a more complex classification method and additional predictors will lead to a higher predictive accuracy and precision of the SPT. The accuracy improves from 0.82 to 0.84 for inhalation allergies, 0.57 to 0.75 for ingestion allergies and 0.69 and 0.80 for all allergies taken together. This improvement in accuracy contributes to a more accurate OFC outcome prediction and may possibly reduce the amount of unnecessary OFC's, thereby reducing the costs and bypassing the discomfort of an OFC for patients and their parents. The paper shows that only a small part of the accuracy improvement derives from a more accurate wheal size determination and that most profit is made from a more complex classifier and using other predictors.

Clinical relevance

Even though paper 2 shows that a more accurate wheal size determination only slightly contributes to a better predictive accuracy, it is relevant to focus on the development of an accurate automated wheal size determination method for the following reasons: First of all, if an accurate and automated wheal size determination is implemented into a smartphone device, general practitioners are able to perform the SPT in secondary care. This will lead to a simple and quick diagnosis, resulting in improvement of the quality of care for the patients. Besides, the implementation will result in time savings for allergists and nurses, especially when the tool is directly linked to the EHR of the patient. Furthermore, an automated wheal size determination method will lead to less intra- and inter-observer variability. This has several benefits: first of all, it will lead to more conformity of SPT results throughout the country, enabling inter-institutional comparison of SPT results. This is important, because this allows multiple institutions to start clinical studies concerning the clinical value of the SPT. An example of such a clinical study, is investigating whether the wheal size of an ingestion allergen (i.e. peanut) can be compensated with a factor times the wheal size of an inhalation allergen (i.e. pollen), in patients with an allergen cross-reaction. The process of allergen cross-reaction means that antibodies produced against inhalation allergens fit on ingestion allergens too. As a result, the wheal size of for example peanut is larger in the presence of a pollen allergy. The effect of cross-reaction results in the fact that ingestion allergies have a poorer predictive accuracy compared to inhalation allergies, as confirmed in paper 2. A wheal segmentation method as proposed in paper 1 could enable a study that investigates an ingestion wheal size compensation in the presence of cross-reaction, which may improve the predictive accuracy of ingestion allergies.

Furthermore, the reduced intra- and inter-observer variability is relevant to follow the sensibilization of a patient for a specific allergen. Patients can overgrow a certain allergy over years. Nowadays, this is assessed by measuring the sIgE level of a patient and it is stated that if the sIgE

level is reduced, it is likely that the patient became tolerant to that allergen. When the SPT is performed in an automated, more accurate and reproducible manner, it could be possible to trace this process by performing a SPT instead of measuring the sIgE level, which is less invasive for the patient.

Even though publication 2 shows that a more complex classifier incorporating multiple clinical factors lead to more correctly classified patients, it is doubtful whether this will lead to a direct reduction in the amount of OFC's. If a patient is classified as being allergic according to the current classifier used in clinic, but as being not allergic according to the new classifier, an accuracy of 75% is not high enough to send patients home for an uncontrolled introduction of the allergen, taken the severity of a possible allergic reaction into account. Furthermore, an accuracy of 75% is often not high enough for patients to avoid the allergen for the rest of their lives, since an allergy is a life-long diagnosis. Besides, patients often want to know what the clinical manifestations of a possible allergy are and may therefore still ask for an OFC.

Recommendations

Future work may focus on further development of the wheal area determination tool. Currently, only the wheal size is determined in pixels, which enables HEWS determination, since the HEWS is a dimensionless outcome measure. However, paper 2 shows that absolute wheal size characteristics contribute to an improved outcome prediction. To include absolute wheal size characteristics, a transformation from pixel size in the image to real world dimensions needs to be made. This can be done with the Aruco marker that is used for perspective correction. However, the antigen drops are placed alternately on the lateral and medial sides of the forearm, since a tape with antigen numbering is located vertically in the center of the forearm. As a result, the wheals are not in the same real world plane and the wheal surfaces are not parallel to the image plane, which makes a translation from pixel size to real world dimensions with one single photo challenging. 3D photography may offer the solution to this. 3D photography can map the arm curvature of a patient and by analyzing the arm curvature, a correction factor can be applied to the wheal area determined with the 2D photo that compensates for the non-parallelism of the wheals to the image plane. 3D photography is evolving rapidly and it may be even possible in the near future to segment the wheal by the small elevation of the wheal in a 3D mesh.

Future work for the diagnostic prediction model could focus on the prediction of ingestion allergies, since this group has the poorest predictive value. This poor predictive value of ingestion allergies can derive from the process of cross-reaction. A solution to overcome the influence of cross-reaction on the prediction of ingestion allergies, is to include the presence of an inhalation allergy as an extra parameter to the the prediction of ingestion allergies.

To finalize, it is relevant to enlarge the datasets for both the studies. For the ResUnet this is required to improve its performance and include more patient variety among which dark skin color. For the prediction model it is required to improve the reliability of the conclusions drawn in this study. Additionally, it may be of interest to investigate the possibilities of an allergen specific prediction model, since different allergens have different predictive characteristics. Within the study, the amount of patients was too low to develop such an antigen specific prediction model. Taken the aim of decentralization into account, an added value would be to focus on the most frequent allergies and develop a prediction model specified on these allergies, for usage at the general practitioner.

7 Conclusion

This thesis presents a (semi-)automatic wheal segmentation method that automates the SPT reading and compared clinical predictors combined in several prediction models, leading to a more accurate, quantitative, objective and reproducible allergy diagnosis. The two papers might enable decentralization from secondary to primary care, allowing inter-institutional comparison of SPT results and improving its diagnostic value.

References

[1] B. I. Nwaru, L. Hickstein, S. Panesar, A. Muraro, T. Werfel, V. Cardona, A. Dubois, S. Halken, K. Hoffmann-Sommergruber, L. K. Poulsen *et al.*, "The epidemiology of food allergy in europe: a systematic review and meta-analysis," *Allergy*, vol. 69, no. 1, pp. 62–75, 2014.

[2] K. Powrie, "Identification and management of drug allergy," *Nursing Standard*, mar 2018.

[3] O. Bulan, "Improved wheal detection from skin prick test images," in *Image Processing: Machine Vision Applications VII*, vol. 9024. International Society for Optics and Photonics, 2014, p. 90240J.

[4] S. Wöhrl, K. Vigl, M. Binder, G. Stingl, and M. Prinz, "Automated measurement of skin prick tests: an advance towards exact calculation of wheal size," *Experimental dermatology*, vol. 15, no. 2, pp. 119–124, 2006.

[5] S. H. Sicherer and H. A. Sampson, "Food allergy: epidemiology, pathogenesis, diagnosis, and treatment," *Journal of Allergy and Clinical Immunology*, vol. 133, no. 2, pp. 291–307, 2014.

[6] J. Van der Valk, R. G. Van Wijk, E. Hoorn, L. Groenendijk, I. M. Groenendijk, and N. De Jong, "Measurement and interpretation of skin prick test results," *Clinical and translational allergy*, vol. 6, no. 1, pp. 1–5, 2015.

[7] R. Vieira dos Santos, R. G. Titus, and H. Cavalcante Lima, "Objective evaluation of skin prick test reactions using digital photography," *Skin Research and Technology*, vol. 13, no. 2, pp. 148–153, 2007.

[8] "IgE in Clinical Allergy and Allergy Diagnosis — World Allergy Organization." [Online]. Available: https://www.worldallergy.org/education-and-programs/education/allergic-disease-resource-center/professionals/ige-in-clinical-allergy-and-allergy-diagnosis

[9] D. J. Unsworth and R. J. Lock, "Food allergy testing," in *Advances in clinical chemistry*. Elsevier, 2014, vol. 65, pp. 173–198.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[13] F. Gao, T. Wu, X. Chu, H. Yoon, Y. Xu, and B. Patel, "Deep residual inception encoder–decoder network for medical imaging synthesis," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 39–49, 2019.

[14] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[15] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[16] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011.

[17] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[18] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[19] "A Gentle Introduction to Random Forests, Ensembles, and Performance Metrics in a Commercial System." [Online]. Available: https://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics

[20] "Feature Selection Using Random forest — by Akash Dubey — Towards Data Science." [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[22] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Information*, vol. 11, no. 6, p. 332, 2020.

[23] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[24] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[25] Z. Zhang, G. Mayer, Y. Dauvilliers, G. Plazzi, F. Pizza, R. Fronczek, J. Santamaria, M. Partinen, S. Overeem, R. Peraita-Adrados *et al.*, "Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[26] "Advantages and Disadvantages of Logistic Regression - GeeksforGeeks." [Online]. Available: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[27] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

[28] "Advantages and Disadvantages of Logistic Regression - GeeksforGeeks." [Online]. Available: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[29] "Mechanism of allergic inflammation in type I hypersensitivity... — Download Scientific Diagram." [Online]. Available: https://www.researchgate.net/figure/Mechanism-of-allergic-inflammation-in-type-I-hypersensitivity-reactions-After-the{_}fig2{_}329559946

Appendices

Appendix A: Sensitization

The presence of atopic diseases begins with the process of sensitization. When the body is exposed to an allergen, the antigens of the allergen are presented to cells involved in the immune response, such as T-lymphocytes. Through a series of specific cell interactions, antibody secretory cells are formed (plasma cells). These plasma cells produce Immunoglobulin E (IgE), which are capable of binding to the specific allergen the body is exposed to. The production of allergen specific IgE-antibodies is known as sensitization. Once formed and released into the circulation, the IgE-antibodies bind on cells such as mast cells and basophils. After this binding, the IgE-antibodies leave their allergen specific receptor available for future interaction with the allergen. When the body is re-exposed to the allergen, the immune system reacts with a more aggressive and rapid memory response. The binding of an allergen with IgE antibodies bound to a mast cell or basophil initiates a process of cell degranulation, with the release of inflammatory mediators. The underlying mechanism of allergy physiology is shown in Figure 8. Sensitization to a certain allergen can manifest itself with a serious allergic reaction i.e. sneezing, an itchy red area and shortness of breath. However, sensitization does not necessarily lead to a clinical allergy and can be asymptomatic. [7–9].

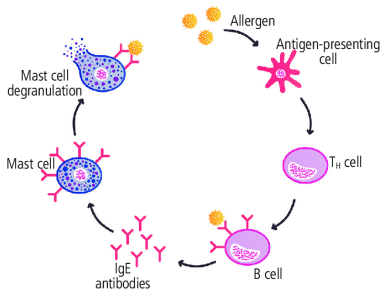


Figure 8. The allergy mechanism [29]

Appendix B: Protocol SPT data collection

Materials

- Apple iPad Pro (2020)
- MacBeth ColorChecker
- 3D printed Aruco marker attached to tourniquet

Procedure

1. Remove iPad cover
2. Check if the iPad settings are correct

Camera application	Settings application
Live Photo OFF	Structuren – Meest compatibel (JPEG/H.264)
HDR ON	Bewaar instellingen – Cameramodus ON Live Photo OFF
Time-lapse / Slowmotion / Video / **Foto** / Portret / Vierkant / Panorama	Raster ON
Slimme HDR ON	
Flash OFF	

3. Ask informed consent to patient and parent
4. Perform SPT according the standard protocol
5. 15 minutes after SPT: Instruct the patient to lay the arm down with hand palm up
6. Locate the Aruco marker tourniquet on the forearm just below the elbow, with the 3D printed arrow alligned with the tape. Place the ColorChecker on left side of patient in length with the forearm
7. Take a picture with the camera as close as possible to the forearm, keeping all wheals and ColorCheker Chart in view. Look at the raster in the camera application to make the photo at 90°. Make sure there is no shading visible over the forearm

Appendix C: Pre-processing

A Python package is used, implementing various colour checker detection algorithms. For installing the Colour - Checker detection software, the following dependencies are required:

- python>=3.5
- colour-science
- opencv-python>=4

Installation of the software is accomplished with the following command:
pip install –user colour-checker-detection
After installation of the software, three main formulas are imported: adjust_image, colour_checkers_coordinates_segmentation and detect_colour_checkers_segmentation.

First, the ColorChecker is segmented from the image. Additionally, a patch of pixels is located in each color swatch, as shown in Figure 9a. The mean pixel value is determined and the color transformation is determined from this value to the reference value, for each swatch. The source and target colors are shown in Figure 9b. To evaluate the result of the transformation, an overlay is made after color correction, which is shown in Figure 9c. The original and color corrected images are shown in Figure 10.

(a) Segmented ColorChecker from image

(b) Overlay extracted colors on reference chart

(c) Overlay result after color correction on reference chart

Figure 9. Color correction algorithm

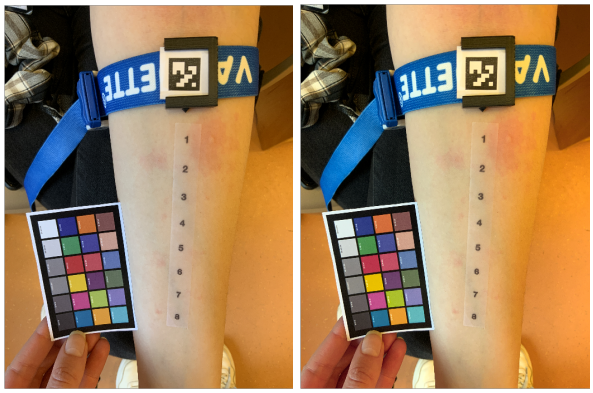


Figure 10. Original image (left) and image after color correction (right)

Appendix D: Light's Kappa

The kappa for each annotator pair is calculated with:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

In which:
P_o = Observed agreement
P_e = Expected agreement

The observed agreement and expected agreement can be calculated as follows:

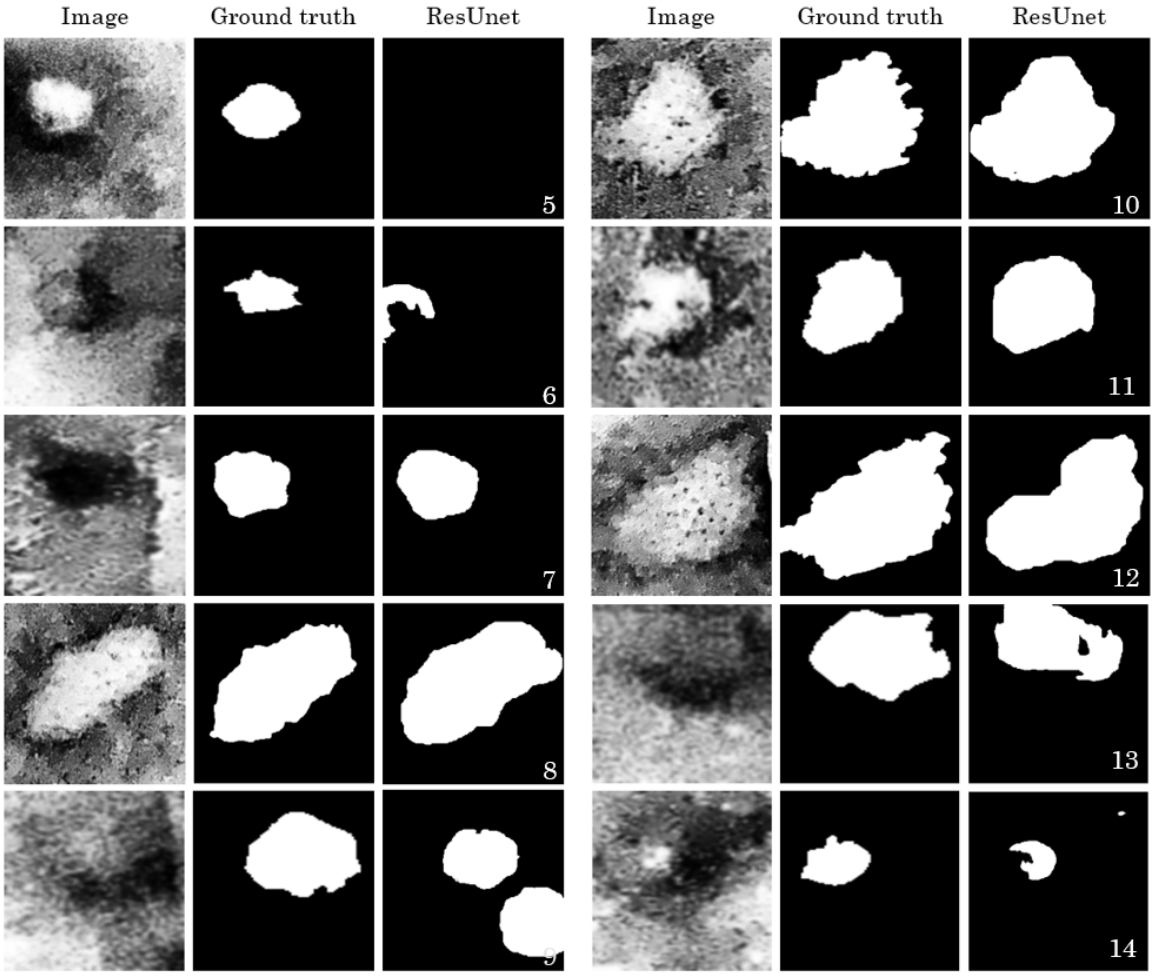
$$P_o = \frac{TP + TN}{n} \tag{7}$$

$$P_e = \frac{\frac{cm^1 rm^1}{n} + \frac{cm^2 rm^2}{n}}{n} \tag{8}$$

In which:
n = total amount of pixels
TP = True Positives. Pixels that are annotated as 1 by both annotators
TN = True Negatives. Pixels that are annotated as 0 by both annotators
cm1 = first column marginal of confusion matrix = TP + FN
cm2 = second column marginal of confusion matrix = FP + TN
rm1 = first row marginal of confusion matrix = TP + FP
rm2 = second row marginal of confusion matrix = FN + TN

Appendix E: Results of ResUnet on test set

This section shows the results of the ResUnet on the test images. The left column shows the input image of the model, the middle column shows the output of the netwerk and the right column shows the ground truth.



Appendix F: Matlab script for wheal segmentation

```
im=imread('174.jpg'); %load jpg of scan

%determine pixelsize
len_a4=297;             %Length of A4 paper in mm
width_a4=210;           %Width of A4 paper in mm
len_im=size(im,1);      %Length of image in pixels
width_im=size(im,2);    %Width of image in pixels
pix_width=width_a4/width_im;    %Mm per pixel
pix_height=len_a4/len_im;       %Mm per pixel
opp_pix=pix_width*pix_height;   %Calculate pixel area


N_boxes= 1:3; %Fill in amount of allergies that are tested
ind_neg=[];   %Empty boxes / negative result
tot_boxes=length(N_boxes);

%Exclude empty boxes and only remain the boxes with wheals
for i = 1:length(ind_neg)
    N_boxes=N_boxes(find(N_boxes~=ind_neg(i)));
end

figure, imshow(im,[])
rect = getrect; %Select the boxes in which allergies are tested
im_cropped=(imcrop(im, [rect]));
im_cropped=rgb2gray(im_cropped);
im_cropped=im2double(im_cropped);

figure, imshow(im_cropped)

delta_x=size(im_cropped,2);             %Width of ROI
delta_y=size(im_cropped,1)/tot_boxes;   %Length of ROI

%Get the ROI rectangles for each box with a wheal in it.
for i = 1:tot_boxes
    crop(i,:)=[1, 1+((i-1)*delta_y), delta_x, delta_y];
    if i==1
        crop(i,:)=[1, 50+((i-1)*delta_y), delta_x, (delta_y-50)];
    end
end
j=0;
figure;
for i = N_boxes
    ROI=(imcrop(im_cropped, crop(i,:)));            %Crop image
    subplot(length(N_boxes),3,1+3*j);imshow(ROI);   %Show wheal

    bg = ROI <= 0.90;               %Create binary mask
    SE=strel('disk',1);             %Create structuring element
    closeBW=imclose(bg,SE);         %Close to create contour of wheal

    filled=imfill(closeBW, 'holes');    %Fill in the contour
    SE=strel('disk',1);                 %Create structuring element
    eroded=imerode(filled,SE);          %Remove the appendices
    dilated=imdilate(eroded,SE);        %Dilate to get original size
    SE=strel('disk',15);                %Remove loose pen stripes
    eroded=imerode(dilated,SE);         %Remove loose pen stripes
    dilated_outer=imdilate(eroded,SE);  %Dilate to get original size
    subplot(length(N_boxes),3,2+(3*j)); imshow(dilated_outer);

    wheal_area(j+1)=sum(dilated_outer(:))*opp_pix;  %Wheal area
    circ_im=bwmorph(dilated_outer, 'remove');       %Contour
    subplot(length(N_boxes),3,(3*j)+3);imshow(circ_im);
    circ(j+1)=sum(circ_im(:))*pix_height;           %Circumference


    j=j+1;
end

hist_area=wheal_area(1);                %Area of positive histamin
hist_circumference=circ(1);             %Circ of positive histamin
for i=1:length(N_boxes)-1
    HEWS(i)=wheal_area(i+1)/hist_area;
    HEWC(i)=circ(i+1)/hist_circumference;
    opp_omt_rat(i)=wheal_area(i+1)/circ(i+1);
end
```

Appendix G: Hyperparameter optimization

Table 2. Table of hyperparameters which are optimized and their ranges

Hyperparameter	**Definition**	**Range**
n_estimators	Number of trees	[30:98:1500]
max_features	Number of features to consider when looking for the best split	['auto', 'sqrt', 'log2']
max_depth	Maximum depth of the three	[2:2:12]
min_samples_split	Minimum number of samples required to split a node	[2:2:20]
min_samples_leaf	Minimum number of samples required to be at a leaf node	[1:1:5]
max_samples	If true, it is the fraction of samples to draw from training data to train each estimator.	[None, 0.7, 0.8, 0.9]
C	Inverse of regularization strength: smaller values specify stronger regularization.	[0.5, 0.7, 1, 1.3, 1.5]
l1_ratio	Used to specify the norm used in the penalization (l1, l2, elasticnet) l1_ratio=0 is equivalent to penalty=l2, l1_ratio=1 is equivalent to penalty=l1, l1_ratio between 0 and 1 is equivalent to penalty = elasticnet	[0, 0.3, 0.5, 0.8, 1]
tolerance	Tolerance for stopping criteria	[5e-5, 7e-5, 1e-4, 3e-4, 5e-4]
fit_intercept	Specifies if a constant should be added to the decision function	[True, False]
max_iter	Maximum number of iterations taken for the solvers to converge	[800,900]

Table 3. Results of 4 fold cross validation for hyperparameter optimization of Random Forest on training set

	All allergies				Inhalation allergies				Ingestion allergies			
Hyperparameter	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4
n_estimators	30	1080	1080	135	30	30	660	1290	870	30	30	30
max_depth	12	6	2	12	12	12	12	2	2	2	2	2
min_samples_split	12	18	10	16	16	14	6	16	12	8	20	4
min_samples_leaf	2	1	4	2	2	3	5	1	1	3	1	4
max_features	'Sqrt'	'log2'	'sqrt'	'log2'	'log2'	'auto'	'auto'	'auto'	'log2'	'auto'	'log2'	'auto'
max_samples	0.9	0.9	0.7	0.7	0.7	0.8	None	0.7	0.7	0.9	0.8	None
Accuracy on Train set	**0.79**	**0.72**	**0.72**	**0.74**	**0.88**	**0.88**	**0.85**	**0.87**	**0.76**	**0.76**	**0.73**	**0.76**
Standard deviation	0.03	0.03	0.04	0.02	0.03	0.03	0.03	0.02	006	0.06	0.08	0.08

Table 4. Results of 4 fold cross validation for hyperparameter optimization of Extra Trees on training set

	All allergies				Inhalation allergies				Ingestion allergies			
Hyperparameter	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4
n_estimators	975	345	1290	30	240	975	1080	135	135	135	555	135
max_depth	8	10	4	12	12	12	8	4	8	12	4	12
min_samples_split	18	10	18	18	4	2	18	18	2	14	6	10
min_samples_leaf	1	10	1	5	2	1	1	1	3	3	2	1
max_features	'log2'	'sqrt'	'sqrt'	'sqrt'	None	'sqrt'	'log2'	'log2'	'log2'	'log2'	'sqrt'	'log2'
max_samples	0.9	None	None	0.7	None	0.7	0.7	0.8	0.7	None	None	None
Accuracy on Train set	**0.80**	**0.74**	**0.72**	**0.73**	**0.86**	**0.87**	**0.85**	**0.87**	**0.76**	**0.76**	**0.74**	**0.74**
Standard deviation	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0.02	0.06	0.07	0.08	0.08

Table 5. Results of 4 fold cross validation for hyperparameter optimization of Gradient Boosting on training set

	All allergies				Inhalation allergies				Ingestion allergies			
Hyperparameter	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4
n_estimators	135	1500	30	30	30	1395	975	30	660	660	30	1080
max_depth	2	4	2	2	2	4	4	6	2	2	2	2
min_samples_split	18	18	12	20	20	20	16	18	16	10	2	6
min_samples_leaf	4	1	5	3	4	5	1	5	3	4	4	2
max_features	'log2'	'auto'	'auto'	'log2'	'sqrt'	'auto'	'log2'	'log2'	'sqrt'	'auto'	'auto'	'log2'
criterion	mse	mae	mse	mae	mse	mae	mae	mae	mae	mae	mae	mae
Accuracy on Train set	**0.78**	**0.72**	**0.72**	**0.73**	**0.89**	**0.88**	**0.85**	**0.87**	**0.84**	**0.75**	**0.72**	**0.75**
Standard deviation	0.01	0.03	0.04	0.02	0.04	0.03	0.02	0.03	0.07	0.08	0.08	0.08

Table 6. Results of 4 fold cross validation for hyperparameter optimization of Logistic Regression on training set

	All allergies				Inhalation allergies				Ingestion allergies			
Hyperparameter	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4	feat 1	feat 2	feat 3	feat 4
C	0.7	1	0.5	0.5	0.7	1.5	0.5	0.5	0.5	0.5	1	0.5
fit_intercept	True	True	True	True	True	False	True	True	True	True	True	True
l1_ratio	0.8	0	0	0	1	0	0	0	0	0	0	0
max_iter	800	900	800	800	900	800	800	800	800	900	800	800
tol	5e-05	5e-05	3e-4	3e-04	5e-05	5e-05	3e-04	3e-04	5e-05	5e-05	5e-05	5e-05
Accuracy on Train set	**0.77**	**0.72**	**0.70**	**0.69**	**0.68**	**0.71**	**0.70**	**0.69**	**0.74**	**0.75**	**0.73**	**0.73**
Standard deviation	0.05	0.02	0.04	0.02	0.04	0.02	0.04	0.02	0.08	0.06	0.07	0.05