Artificial Intelligence in Dry Eye Disease

Andrea M. Storås[a,e], Inga Strümke[a], Michael A. Riegler[a], Jakob Grauslund[b,c,d], Hugo L. Hammer[a,e], Anis Yazidi[e], Pål Halvorsen[a,e], Kjell G. Gundersen[h], Tor P. Utheim[e,f,g], Catherine J. Jackson[h]

[a] *SimulaMet, Oslo, Norway*
[b] *Department of Ophthalmology, Odense University Hospital, Odense, Denmark*
[c] *Department of Clinical Research, University of Southern Denmark, Odense, Denmark*
[d] *Department of Ophthalmology, Vestfold University Trust, Tønsberg, Norway*
[e] *Department of Computer Science, Oslo Metropolitan University, Norway*
[f] *Department of Medical Biochemistry, Oslo University Hospital, Norway*
[g] *Department of Ophthalmology, Oslo University Hospital, Norway*
[h] *Ifocus, Haugesund, Norway*

Abstract

Dry eye disease (DED) has a prevalence of between 5 and 50%, depending on the diagnostic criteria used and population under study. However, it remains one of the most underdiagnosed and undertreated conditions in ophthalmology. Many tests used in the diagnosis of DED rely on an experienced observer for image interpretation, which may be considered subjective and result in variation in diagnosis. Since artificial intelligence (AI) systems are capable of advanced problem solving, use of such techniques could lead to more objective diagnosis. Although the term 'AI' is commonly used, recent success in its applications to medicine is mainly due to advancements in the sub-field of machine learning, which has been used to automatically classify images and predict medical outcomes. Powerful machine learning techniques have been harnessed to understand nuances in patient data and medical images, aiming for consistent diagnosis and stratification of disease severity. This is the first literature review on the use of AI in DED. We provide a brief introduction to AI, report its current use in DED research and its potential for application in the clinic. Our review found that AI has been employed in a wide range of DED clinical tests and research applications, primarily for interpretation of interferometry, slit-lamp and meibography images. While initial results are promising, much work is still needed on model development, clinical testing and standardisation.

Keywords: dry eye disease, artificial intelligence, machine learning

1. Introduction

Dry eye disease (DED) is one of the most common eye diseases worldwide, with a prevalence of between 5 and 50%, depending on the diagnostic criteria used and study population [1]. Yet, although symptoms stemming from DED are reported as the most common reason to seek medical eye care [1], it is considered one

*Corresponding author. SimulaMet, Oslo, Norway
Email address: andrea@simula.no (Andrea M. Storås)

of the most underdiagnosed and undertreated conditions in ophthalmology [2]. Symptoms of DED include eye irritation, photophobia and fluctuating vision. The condition can be painful and might result in lasting damage to the cornea through irritation of the ocular surface. Epidemiological studies indicate that DED is most prevalent in women [3] and increases with age [1]. However, the incidence of DED is likely to increase in all age groups in coming years due to longer screen time and more prevalent use of contact lenses, which are both risk factors [4]. Other risk factors include diabetes mellitus [5] and exposure to air-pollution [6]. DED can have a substantial effect on the quality of life, and may impose significant direct and indirect public health costs as well as personal economic burden due to reduced work productivity.

DED is divided into two subtypes defined by the underlying mechanism of the disease: (i) aqueous deficient DED, where tear production from the lacrimal gland is insufficient and (ii) evaporative DED (the most common form), which is typically caused by dysfunctional meibomian glands in the eyelids. Meibomian glands are responsible for supplying meibum, which is a concentrated substance that normally covers the surface of the cornea to form a protective superficial lipid layer that guards against evaporation of the underlying tear film. The ability to reliably distinguish between aqueous deficient and evaporative DED, their respective severity levels and mixed aqueous/evaporative forms is important in deciding the ideal modality of treatment. A fast and accurate diagnosis relieves patient discomfort and also spares them unnecessary expense and exposure to potential side effects associated with some treatments. A tailor made treatment plan can yield improved treatment response and maximize health provider efficiency.

The main clinical signs of DED are decreased tear volume, more rapid break-up of the tear film (fluorescein tear break-up time (TBUT)) and microwounds of the ocular surface [7]. In the healthy eye, the tear film naturally 'breaks up' after ten seconds and the protective tear film is reformed with blinking. Available diagnostic tests often do not correlate with the severity of clinical symptoms reported by the patient. No single clinical test is considered definitive in the diagnosis of DED[1]. Therefore, multiple tests are typically used in combination and supplemented by information gathered on patient symptoms, recorded through questionnaires. These tests demand a significant amount of time and resources at the clinic. Tests for determining the physical parameters of tears include TBUT, the Schirmer's test, tear osmolarity and tear meniscus height. Other useful tests in DED diagnosis include ocular surface staining, corneal sensibility, interblink frequency, corneal surface topography, interferometry, aberrometry and imaging techniques such as meibography and in vivo confocal microscopy (IVCM), as well as visual function tests.

Artificial intelligence (AI) was defined in 1955 as "the science and engineering of making intelligent machines" [8], where intelligence is the "ability to achieve goals in a wide range of environments" [9]. Within AI, machine learning denotes a class of algorithms capable of learning from data rather than being programmed with explicit rules. AI, and particularly machine learning, is increasingly becoming an integral part of health

care systems. The sub-field of machine learning known as deep learning uses deep artificial neural networks, and has gained increased attention in recent years, especially for its image and text recognition abilities. In the field of ophthalmology, deep learning has so far mainly been used in the analysis of data from the retina to segment regions of interest in images, automate diagnosis and predict disease outcomes [10]. For instance, the combination of deep learning and optical coherence tomography (OCT) technologies has allowed reliable detection of retinal diseases and improved diagnosis [11]. Machine learning also has potential for use in the diagnosis and treatment of anterior segment diseases, such as DED and has already found its way into the field with methods such as presented by Ciezar et al. [12]. Many of the tests used for DED diagnosis and follow-up rely on the experience of the observer for interpretation of images, which may be considered subjective [13]. AI tools can be used to interpret images automatically and objectively, saving time and providing consistency in diagnosis.

Several reviews have been published that discuss the application of AI in eye disease, including screening for diabetic retinopathy [14], detection of age-related macular degeneration [15] and diagnosis of retinopathy of prematurity [16]. We are, however, not aware of any review on AI in DED. In this article, we therefore provide a critical review of the use of AI systems developed within the field of DED, discuss their current use and highlight future work.

2. Artificial intelligence

AI is informational technology capable of performing activities that require intelligence. It has gained substantial popularity within the field of medicine due to its ability to solve ubiquitous medical problems, such as classification of skin cancer [17], prediction of hypoxemia during surgeries [18], identification of diabetic retinopathy [19] and prediction of risk for future need of keratoplasty [20]. Machine learning is a sub-field of AI encompassing algorithms capable of learning from data, without being explicitly programmed. All AI systems used in the studies included in this review, fall within the class of machine learning. The process by which a machine learning algorithm learns from data is referred to as *training*. The outcome of the training process is a machine learning *model*, and the model's output is referred to as *prediction*s. Different learning algorithms are categorised according to the type of data they use, and referred to as supervised, unsupervised and reinforcement learning. The latter is excluded from this review, as none of the studies use it, while the two former are introduced in this section. A complete overview of the algorithms encountered in the reviewed studies is provided in Figure 1, sorted according to the categories described below.

2.1. Supervised learning

Supervised learning denotes the learning process of an algorithm using labelled data, meaning data that contains the target value for each data instance, e.g., tear film lipid layer category. The learning process involves extracting patterns linking the input variables and the target outcome. The performance of the resulting model is evaluated by letting it predict on a previously unseen data set, and comparing the predictions to the true data labels. See Section 2.5 for a brief discussion of evaluation metrics. Supervised learning algorithms can perform regression and classification, where regression involves predicting a numerical value for a data instance, and classification involves assigning data instances to predefined categories. Figure 1 contains an overview of supervised learning algorithms encountered in the reviewed studies.

2.2. Unsupervised learning

Unsupervised learning denotes the training process of an algorithm using unlabelled data, i.e., data not containing target values. The task of the learning algorithm is to find patterns or data groupings by constructing a compact representation of the data. This type of machine learning is commonly used for grouping observations together, detecting relationships between input variables, and for dimensionality reduction. As unsupervised learning data contains no labels, a measure of model performance depends on considerations outside the data [see 21, chap. 14], e.g., how the task would have been solved by someone in the real world. For clustering algorithms, similarity or dissimilarity measures such as the distance between cluster points can be used to measure performance, but whether this is relevant depends on the task [22]. Unsupervised algorithms encountered in the reviewed studies can be divided into those performing clustering and those used for dimensionality reduction, see Figure 1 for an overview.

2.3. Artificial neural networks and deep learning

Artificial neural networks are loosely inspired by the neurological networks in the biological brain, and consist of artificial neurons organised in layers. How the layers are organised within the network is referred to as its *architecture*. Artificial neural networks have one input layer, responsible for passing the data to the network, and one or more hidden layers. Networks with more than one hidden layer are called deep neural networks. The final layer is the output layer, providing the output of the entire network. Deep learning is a sub-field of machine learning involving training deep neural networks, which can be done both in a supervised and unsupervised manner. We encounter several deep architectures in the reviewed studies. The two more advanced types are convolutional neural networks (CNNs) and generative adversarial networks (GANs). CNN denotes the commonly used architecture for image analysis and object detection problems, named for having so-called convolutional layers that act as filters identifying relevant features in images. CNNs have

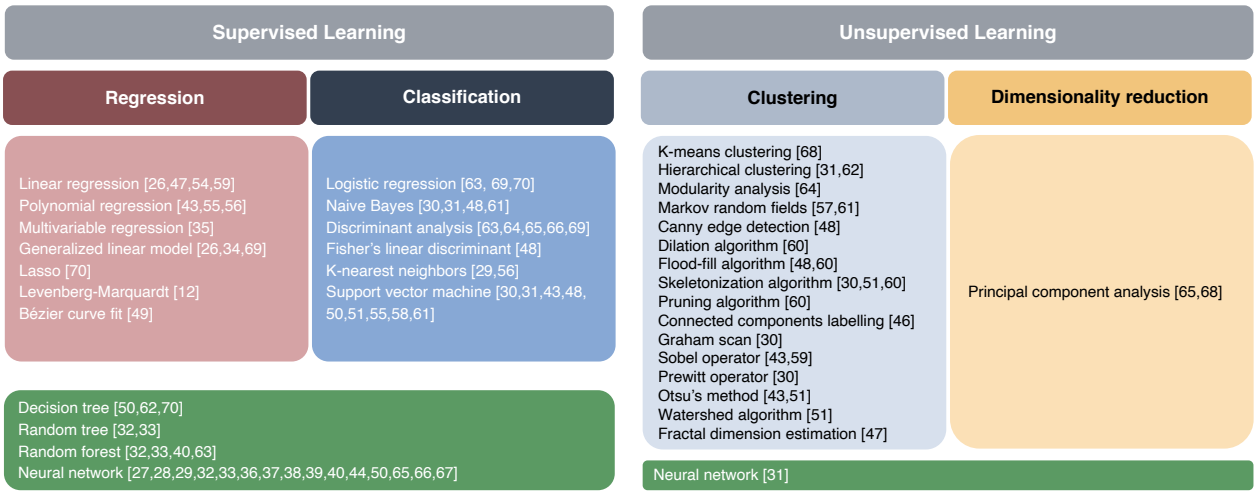


Figure 1: An overview of the machine learning algorithms used in the reviewed studies.

gained popularity recently and all of the reviewed studies that apply CNNs were published in 2019 or later. Advanced deep learning techniques will most likely replace the established image analysis methods. This trend has been observed within other medical fields such as gastrointestinal diseases and radiology [23, 24]. A GAN is a combination of two neural networks: A generator and a discriminator competing against each other. The goal of the generator is to produce fake data similar to a set of real data. The discriminator receives both real data and the fake data from the generator, and its goal is to discriminate the two. GANs can among other things be used to generate synthetic medical data, alleviating privacy concerns [25].

2.4. Workflow for model development and validation

The data used for developing machine learning models is ideally divided into three independent parts: A training set, a validation set and a test set. The training set is used to tune the model, the validation set to evaluate performance during training, and the test set to evaluate the final model. A more advanced form of training and validation, is k-fold cross-validation. Here, the data is split into k parts, of which one part is set aside for validation, while the model is trained on the remaining data. This is repeated k times, and each time a different part of the data is used for validation. The model performance can be calculated as the average performance for the k different models [see 21, chap. 7]. It is considered good practice to not use the test data during model development and vice versa, the model should not be tuned further once it has been evaluated on the test data [see 21, chap.7]. In cases of class imbalance, i.e., unequal number of instances

from the different classes, there is a risk of developing a model that favors the prevalent class. If the data is stratified for training and testing, this might not be captured during testing. Class imbalance is common in medical data sets, as there are for instance usually more healthy than ill people in the population [26]. Whether to choose a class distribution that represents the population, a balanced or some other distribution depends on the objective. Various performance scores should regardless always be used to provide a full picture of the model's performance.

2.5. Performance scores

In order to assess how well a machine learning model performs, its performance can be assigned a score. In supervised learning, this is based on the model's output compared to the desired output. Here, we introduce scores used most frequently in the reviewed studies. Their definitions as well as the remaining scores used are provided in Appendix A.1. A commonly used performance score in classification is *accuracy*, Equation (A.3), which denotes the proportion of correctly predicted instances. Its use is inappropriate in cases of strong class imbalance, as it can reach high values if the model always predicts the prevalent class. The *sensitivity*, also known as recall, Equation (A.4), denotes the true positive rate. If the goal is to detect all positive instances, a high sensitivity indicates success. The *precision*, Equation (A.5), denotes the positive predictive value. The *specificity*, Equation (A.6), denotes the true negative rate, and is the negative class version of the sensitivity. The *F1 score*, Equation (A.7), is the harmonic mean between the sensitivity and the precision. It is not symmetric between the classes, meaning it is dependent on which class is defined as positive.

Image segmentation involves partitioning the pixels in an image into segments [27]. This can for example be used to place all pixels representing the pupil into the same segment while pixels representing the iris are placed in another segment. The identified segments can then be compared to manual annotations. Performance scores used include the *Average Pompeiu-Hausdorff distance*, (A.17), the *Jaccard index* and the *support*, all described in Appendix A.1.

2.6. AI regulation

Approved AI devices will be a major part of the medical service landscape in the future. Currently, many countries are actively working on releasing AI regulations for healthcare, including the European Union (EU), the United States, China, South Korea and Japan. On 21 April 2021, the EU released a proposal for a regulatory framework for AI [28]. The US Food and Drug Administration (FDA) is also working on AI legislation for healthcare [29].

In the framework proposed by the EU, AI systems are divided into the four categories low risk, minimal risk, high risk and unacceptable risk [28]. AI systems that fall into the high risk category are expected to

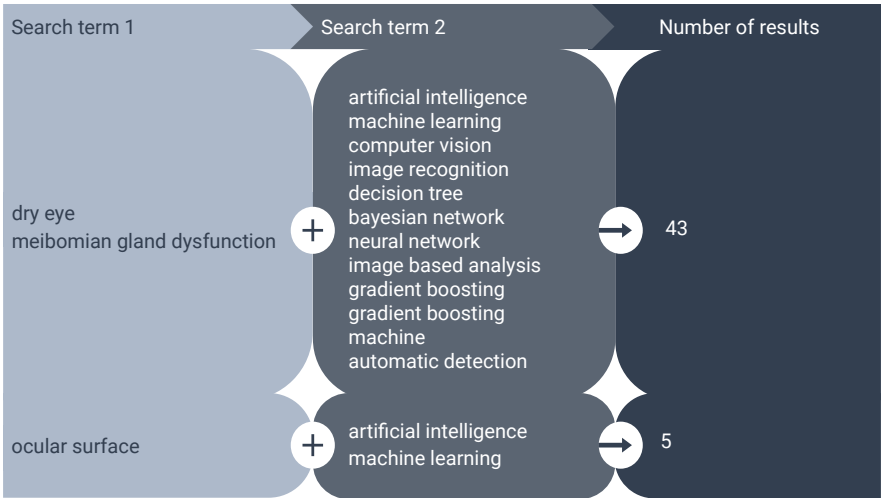


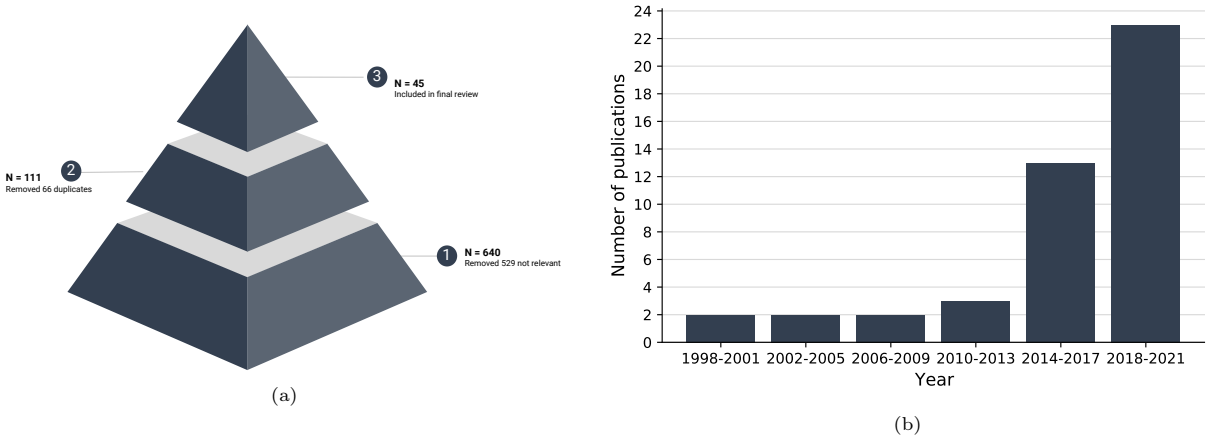
Figure 2: Search term combinations used in the literature search. Three of the studies found in the searches including "ocular surface" were also found among the studies in the searches including "dry eye".

be subject to strict requirements, including data governance, technical documentation, transparency and provision of information to users, human oversight, robustness and cyber security, and accuracy. It is highly likely that medical devices using AI will end up in the high risk category. Looking at the legislation proposals [28, 29] from an AI research perspective, it is clear that explainable AI, transparency, uncertainty assessment, robustness against adversarial attacks, high quality of data sets, proper performance assessment, continuous post-deployment monitoring, human oversight and interaction between AI systems and humans, will be major research topics for the development of AI in healthcare.

3. Methods

3.1. Search methods

A systematic literature search was performed in PubMed and Embase in the period between March 20 and May 21, 2021. The goal was to retrieve as many studies as possible applying machine learning to DED related data. The following keywords were used: All combinations of "dry eye" and "meibomian gland dysfunction" with "artificial intelligence", "machine learning", "computer vision", "image recognition", "bayesian network", "decision tree", "neural network", "image based analysis", "gradient boosting", "gradient boosting machine" and "automatic detection". In addition, searches for "ocular surface" combined with both "artificial intelligence" and "machine learning" were made. See also an overview of the search terms and combinations in Figure 2. No time period limitations were applied for any of the searches.



(a) (b)

Figure 3: (a) Illustration of the three steps in the study selection process and number of studies (**N**) included in each step, and (b) the number of studies published over time, counting the studies included in this review.

3.2. Selection criteria

The studies to include in the review had to be available in English in full-text. Studies not investigating the medical aspects of DED were excluded (e.g., other ocular diseases and cost analyses of DED). Moreover, the studies had to describe the use of a machine learning model in order to be considered. Reviews were not considered. The studies were selected in a three-step process. One review author screened the titles on the basis of the inclusion criteria. The full-texts were then retrieved and studied for relevance. The search gave 640 studies in total, of which 111 were regarded as relevant according to the selection criteria. After removing duplicates, 45 studies were left. The three-step process is shown in Figure 3a.

4. Artificial intelligence in dry eye disease

4.1. Summary of the studies

Most studies were published in recent years, especially after 2014, see Figure 3b. An overview of the studies is provided in Tables 1 to 4 for the clinical, biochemical and demographical studies, respectively. Information on the data used in each study is shown in Table 5. We grouped studies according to the type of clinical test or type of study: TBUT, interferometry and slit-lamp images, IVCM, meibography, tear osmolarity, proteomics analysis, OCT, population surveys and other clinical tests. We found most studies employed machine learning for interpretation of interferometry, slit-lamp and meibography images.

Table 1: Overview of the reviewed studies using clinical investigations, part 1 of 2.

Study	Objective	N	Clinical Tests	Type of Data	Type of Algorithm	Performance Score(s)
Aggarwal S et al. (2021) [30]	DED mechanism, effect of therapy	199	Subjective symptoms, Schirmer´s test with anasthesia, TBUT, vital staining of cornea and conjunctiva, laser IVCM images, subbasal layer of cornea: DC density and morphology	Images of cornea	GLM, MLR	GLM: p-values < 0.05 for DC density and number of DCs, MLR: p-values < 0.05 between DC density and CFS, number of DCs and CFS, DC size and CFS, DC density and conjunctival staining, number of DCs and TBUT, corresponding β-coefficients = 0.20, -0.23, 0.36, 0.24 and -0.18
Deng X et al. (2021) [31]	Estimate tear meniscus height	217	Oculus Keratograph	Tear meniscus images	CNN (U-net)	Accuracy = 82.5%, sensitivity = 0.899, precision = 0.911, F1 score = 0.901
Elsawy A et al. (2021) [32]	Diagnose DED	547	AS-OCT	Ocular surface images	Pretrained CNN (VGG19)	AUCROC = 0.99 (model 1) and 0.98 (model 2), AUCPRC = 0.96 (model 1) and 0.94 (model 2), F1 score = 0.90 (model 1) and 0.86 (model 2)*
Khan ZK et al. (2021)[33]	Detect MGD	112	Meibomian gland 3D IR-images, lower and upper eyelid	Meibomian gland images	GAN	F1 score = 0.825, P-HD = 4.611, aggregated JI = 0.664, r = 0.962 (clincian1) and 0.968 (clinician2), p-values < 0.001, mean difference = 0.96 (clincian1) and 0.95 (clincian2)
Xiao P et al.(2021) [34]	Detect MGD	15 (images)	Oculus Keratograph	IR meibography images	Prewitt operator, Graham scan algorithm, fragmentation algorithm and SA (used sequentially)	Gland area: KI = 0.94, FPR = 6.02%, FNR = 6.43%. Gland segmentation: KI = 0.87, FPR = 4.35%, FNR = 18.61%*
Yeh C-H et al. (2021) [35]	Detect MGD	706 (images)	Oculus Keratograph	IR meibography images	Nonparametric instance discrimination, pretrained CNN (ImageNet), hierarchical clustering	: Accuracy: meiboscore grading = 80.9%, 2-class classification = 85.2%, 3-class classification = 81.3%, 4-class classification = 80.8%*
da Cruz LB et al. (2020)[36]	Classify tear film patterns	106 (images)	Doane interferometer	Tear film lipid layer images	SVM, RF, RT, Naive Bayes, DNN, simple NN	RF: accuracy = 97.54%, SD = 0.51%, F1 score = 0.97, KI = 0.96, AUCROC = 0.99**
da Cruz LB et al. (2020) [37]	Classify tear film patterns	106 (images)	Doane interferometer	Tear film lipid layer images	SVM, RF, RT, Naive Bayes, DNN, simple NN	RF: accuracy = 99.622%, SD = 0.843%, F1 score = 0.996, KI = 0.995, AUCROC = 0.999***
Fu P-I et al. (2020) [38]	Compare 2 methods	28	Oculus Keratograph	Tear film lipid layer images (with and without preprocessing)	GLM	β-coefficients = 0.6, 10
Fujimoto K et al. (2020) [39]	Compare 2 methods	195	Pentacam vs AS-OCT	CCT, TCT, thinnest point of cornea	Multivariable regression	Severe DED: β-coefficients = 7.029 (CCT) and 6.958 (TCT), p-values = 0.002 (CCT) and 0.049 (TCT), 95% CI = 2.528 − 11.530 (CCT) and 0.037 − 13.879 (TCT)
Maruoka S et al. (2020) [40]	Detect MGD	221	IVCM	Meibomian gland images	Combinations of 9 CNNs	Single CNN: AUROC = 0.966, sensitivity = 0.942, specificity = 0.821, ensemble CNNs: AUROC = 0.981, sensitivity = 0.921, specificity = 0.988
Prabhu SM et al. (2020) [41]	Quantify and detect MGD	400 (images)	Oculus Keratograph, digital camera	CNN (U-net)	p-values > 0.005 between model output and clinical experts	
Stegmann H et al. (2020) [42]	Detect tear meniscus in images	10	Optical coherence tomography	Tear meniscus images	2 CNNs	Meniscus localization: JI = 0.7885, sensitivity = 0.9999, meniscus segmentation best CNN: accuracy = 0.9995, sensitivity = 0.9636 , specificity = 0.9998, JI = 0.9324, F1 score = 0.9644, support = 0.0071*, ***
Wei S et al. (2020) [43]	DED mechanism, effect of therapy	53	Corneal IVCM with anesthesia	Images of cornea	Pretrained CNN (U-net)	AUROC = 0.96, sensitivity = 96%
Giannaccare G et al. (2019) [44]	Subbasal nerve plexus characteristics for diagnosing DED	69	IVCM	Images of subbasal nerve plexus	Earlier developed method involving RF and NN [45, 46]	Nan

Abbreviations: N = number of subjects; DED = dry eye disease; IVCM = in vivo confocal microscopy; DC = dendritic cell; GLM = generalized linear model; MLR = multiple linear regression; CFS = corneal fluorescein score; AS-OCT = anterior segment optical coherence tomography; CNN = convolutional neural network; AUROC = area under reciever operating characteristic curve; AUPRC = area under precision-recall curve; MGD = meibomian gland dysfunction; GAN = generative adversarial network; P-HD = average Pompeiu-Hausdorff distance; JI = Jaccard index; KI = Kappa index; CTRL = healthy; FPR = false positive rate; FNR = false negative rate; SVM = support vector machine; RF = random forest; RT = random tree; DNN = deep neural network; SD = standard deviation; CCT = central corneal thickness; TCT = thinnest corneal thickness; r = Pearson's correlation coefficient;Nan = not available; NN = neural network; RMSE = root mean squared error; CI = confidence interval; TBUT = fluorescein tear break-up time; PA = pruning algorithm; SA = skeletonization algorithm; FFA = Flood-fill algorithm; * = standard deviations not included in table; ** = 95% confidence intervals not included in table; *** = metrics are calculated as the average of 5 repetitions;**** = metrics are calculated as the average of 10 repetitions; ***** = metrics are calculated as the average from 10-fold crossvalidation; $\ominus$ = metrics are calculated as the average from 6-fold crossvalidation $\oplus$ = metrics are calculated as the average of 100 models)

Table 2: Overview of the reviewed studies using clinical investigations, part 2 of 2.

Study	Objective	N	Clinical Tests	Type of Data	Type of Algorithm	Performance Score(s)
Llorens-Quintana C et al. (2019) [47]	Evaluate meibomian gland atrophy	149	Oculus Keratograph	Meibography images	Sobel operator, polynomial function, fragmentation algorithm, Otsu's method (used sequentially)	p-values < 0.05 between automatic method and clinicians
Wang J et al. (2019) [48]	Evaluate meibomian gland atrophy	706 (images)	Oculus Keratograph	Meibography images	4 CNNs	Meiboscore grading: accuracy = 95.6%, eyelid detection: accuracy = 97.6%, JI = 0.955, atrophy detection: accuracy = 95.4%, JI = 0.667, RMSE = 0.067 (average across 4 meiboscores)
Yabusaki K (2019) [49]	Diagnose DED	138 (images)	Tear interferometer	Tear film lipid layer images	SVM	KI = 0.820, CTRL: F1 score = 0.845, SD = 0.067, aqueous-deficient DED: F1 score = 0.981, SD = 0.023, evaporative DED: F1 score = 0.815, SD = 0.095****
Yang J et al. (2019) [50]	Estimate tear meniscus height for DED	69	Slit-lamp images with fluorescence staining	Ocular surface images	Connected component labelling	Mean: p-value < 0.01 (x16 and x40 magnification), r = 0.626 (x16) and 0.711 (x40), max: p-value < 0.001 (x16 and x40), r = 0.645 (x16) and 0.847 (x40)
Szyperski PD (2018) [51]	Diagnose DED	110	Interferometry	Videos from lateral shearing interferometry	4 different fractal dimension estimators, linear regression	Best estimator: AUROC = 0.786
Hwang H et al. (2017) [52]	Estimate tear film lipid layer thickness	34	Lipiscanner 1.0, slit-lamp microscope	Tear film lipid layer videos	Flood-fill algorithm, Canny edge detection	p-value < 0.01 between all MGD groups
Koprowski R et al. (2017) [53]	Detect MGD	57	Oculus Keratograph	Meibography images	Riesz pyramid (?), Bezier curve (used sequentially)	Accuracy = 99.08%, sensitivity = 1, specificity = 0.98
Peteiro-Barral D et al. (2017) [54]	Classify tear film patterns	105 (images)	Tearscope plus images	Tear film lipid layer images	SVM, Decision tree, Naive Bayes, simple NN, Fisher´s linear discriminant	NN: accuracy = 96%, sensitivity = 92%, specificity = 97%, precision = 92%, F1 score = 0.93, AUCROC = 0.95
Koprowski et al. (2016) [55]	Detect MGD	86	Oculus Keratograph	Meibography images	Otsu's method, SA, watershed algorithm (used sequentially)	Sensitivity = 0.993, specificity = 0.975
Remeseiro B et al. (2016) [56]	Classify tear film patterns	128 (images)	Tearscope-plus images	Tear film lipid layer images	SVM	accuracy = 96.09%, precision = 92.00%, sensitivity = 89.66%, specificity = 97.98%, F1 score = 91.23%, processing time = 0.07 s
Remeseiro B et al. (2016) [57]	Classify tear film patterns	50 (images)	Tearscope-plus images	Tear film lipid layer images	SVM	accuracy = 90.89%, sensitivity = 83.54%, precision = 97.95%, specificity = 86.75%
Kanellopoulos AJ et al. (2014) [58]	Diagnose DED	70	Fourier-domain AS-OCT system: corneal and corneal epithelial thickness maps	Corneal examination	Linear regression (correlation between DED and thickness)	Nan
Ramos L et al. (2014) [59]	Estimate TBUT	18 (videos)	Videos from TBUT (slit-lamp)	TBUT videos	Polynomial function	specificity = 89% (parameter b) and 82% (parameter e), specificity = 84% and 80%
Ramos L et al. (2014) [60]	Estimate TBUT	18 (videos)	Videos from TBUT (slit-lamp)	TBUT videos	Polynomial function	accuracy = "more than 90%"
Remeseiro et al. (2014) [61]	Classify tear film patterns	511 (images)	Tearscope-plus images	Tear film lipid layer images	Markov random field, SVM (used sequentially)	accuracy = 97.14%, accuracy (noisy data) = 92.61%*****
García-Resúa C et al. (2013) [62]	Classify tear film patterns	105	Tearscope-plus images	Tear film lipid layer images	K-nearest neighbors	Cramér's V = 0.9, r = 0.94, p-value < 0.001, accuracy = 86.2% $\ominus$
Rodriquez JD (2013) [63]	Evaluate ocular redness	26	Slit-lamp, digital camera	Images of conjunctiva	Sobel operator, MLR (used sequentially)	Accuracy = 100%, r = 0.76, concordance correlation = 0.76 (compared to 5 investigators)*
Koh YW et al. (2012) [64]	Detect MGD	55	Slit-lamp biomicroscope, upper eye lid	IR meibography images	PA, SA, FFA, SVM (used sequentially)	specificity = 96.1%, SD = 0.4%, sensitivity = 97.9%, SD = 0.6% $\oplus$
Yedidya T et al. (2009) [65]	Estimate TBUT	22 (videos)	Video from TBUT	TBUT videos	Markov random field	average difference in TBUT = 2.34s
Yedidya T et al. (2007) [13]	Detect dry areas	8 (videos)	Video from TBUT	TBUT videos	Levenberg-Marquardt	accuracy = 91% (84 − 96%), SD = 4%
Mathers WD et al. (2004) [66]	Investigate DED	513	Schirmer´s test, meibomian gland drop-out, lipid viscosity and volume, tear evaporation	Clinical test results	Hierarchical clustering, decision tree	Nan

Abbreviations: N = number of subjects; DED = dry eye disease; IVCM = in vivo confocal microscopy; DC = dendritic cell; GLM = generalized linear model; MLR = multiple linear regression; CFS = corneal fluorescein score; AS-OCT = anterior segment optical coherence tomography; CNN = convolutional neural network; AUROC = area under reciever operating characteristic curve; AUPRC = area under precision-recall curve; MGD = meibomian gland dysfunction; GAN = generative adversarial network; P-HD = average Pompeiu-Hausdorff distance; JI = Jaccard index; KI = Kappa index; CTRL = healthy; FPR = false positive rate; FNR = false negative rate; SVM = support vector machine; RF = random forest; RT = random tree; DNN = deep neural network; SD = standard deviation; CCT = central corneal thickness; TCT = thinnest corneal thickness; r = Pearson's correlation coefficient;Nan = not available; NN = neural network; RMSE = root mean squared error; CI = confidence interval; TBUT = fluorescein tear break-up time; PA = pruning algorithm; SA = skeletonization algorithm; FFA = Flood-fill algorithm; * = standard deviations not included in table; ** = 95% confidence intervals not included in table; *** = metrics are calculated as the average of 5 repetitions;**** = metrics are calculated as the average of 10 repetitions; ***** = metrics are calculated as the average from 10-fold crossvalidation; $\ominus$ = metrics are calculated as the average from 6-fold crossvalidation $\oplus$ = metrics are calculated as the average of 100 models)

Table 3: Overview of the reviewed studies using biochemical investigations.

Study	Objective	N	Clinical Tests	Type of Data	Type of Algorithm	Performance Score(s)
Cartes C et al. (2019) [67]	Diagnose DED	40	Tear-Lab Osmometer	Tear osmolarity measurements	LR, Naive Bayes, SVM, RF	LR: accuracy = 85%
Jung JH et al. (2017) [68]	Detect protein patterns in DED	10	Pooled tear and lacrimal fluid, analysed with LC-MS, trypsin digestion, RP-LC fractionation	Proteins in tears and lacrimal fluid	"Network model" based on betweenness centrality	Nan
Gonzalez N (2014) [69]	Diagnose DED	93	Peptide/protein analysis: gel electrophoresis (SDS-PAGE)	Peptides and proteins in tears	Discriminant analysis, PCA, NN	Accuracy = 89.3%, CTRL: sensitivity = 0.99, specificity = 0.96, MGD: sensitivity = 0.85, specificity = 0.96, aqueous-deficient DED: sensitivity = 0.83, specificity = 0.93*
Grus FH et al. (2005) [70]	Diagnose DED	159	Schirmer´s test with anesthesia, tears analysed by LC-MS	Proteins in tears	Discriminant analysis, DNN (used sequentially)	AUROC = 0.93, sensitivity and specificity = "approx. 90% each"
Grus FH et al. (1999) [71]	Diagnose DED	60	Protein analysis: gel electrophoresis (SDS-PAGE)	Proteins in tears	DNN, discriminant analysis	DNN: accuracy = 89%, discriminant analysis: accuracy = 71%
Grus FH et al. (1998) [72]	Diagnose DED	119	Protein analysis: gel electrophoresis (SDS-PAGE)	Proteins in tears	Principal component analysis, K-means clustering (used sequentially), discriminant analysis	K-means: accuracy = 71% (DED vs CTRL) and 42% (DED, diabetes-DED, CTRL), discriminant analysis: accuracy = 72% (DED vs CTRL) and 43% (DED, diabetes-DED, CTRL)

Abbreviations: N = number of subjects; DED = dry eye disease; LR = logistic regression; SVM = support vector machine; RF = random forest; AUROC = area under reciever operating characteristic curve; MGD = meibomian gland dysfunction; CTRL = healthy; DNN = deep neural network; Nan = not available; NN = neural network; LC-MS = liquid chromotography mass spectometry; RP-LC = reverse-phase liquid chromatography; SDS-PAGE = sodium dodecyl sulphate-polyacrylamide gel electrophoresis; OSDI = ocular surface disease index; * = metrics are calculated as the average of 10 repetitions

Table 4: Overview of the reviewed studies using demographical investigations.

Study	Objective	N	Clinical Tests	Type of Data	Type of Algorithm	Performance Score(s)
Choi HR et al. (2020) [73]	Investigate DED and dyslipidemia association	2272	OSDI score, health examination, questionnaire	Population studies, Korea	GLM, LR	Nan
Nam SM et al. (2020) [74]	Detect risk factors for DED	4391	Health examination, health survey, nutrition survey	National health survey, Korea	Decision tree, Lasso, LR (used sequentially)	AUROC = 0.70, 95% CI = 0.61 − 0.78, specificity = 68%, sensitivity = 66%
Kaido M et al. (2015) [75]	Diagnose DED	369	Blink frequency, visual maintenance ratio, questionnaire	Functional VA measurement and questionnaire, Japanese visual display terminal workers	Discriminant analysis	sensitivity = 93.1%, specificity = 43.7%, precision = 83.8%, NPV = 80.8%

Abbreviations: N = number of subjects; DED = dry eye disease; GLM = generalized linear model; AUROC = area under reciever operating characteristic curve; Nan = not available; CI = confidence interval; LR = logistic regression; OSDI = ocular surface disease index; VA = visual acuity; NPV = negative predictive value

Table 5: Overview of the data applied for the analyses.

Study	Type of Input Data	Training Dataset	Testing Dataset	Reference Standard
Clinical Investigations				
Aggarwal S et al. (2021) [30]	Tabular	349	Nan	Nan (clinical test results, subjective report)
Deng X (2021) [31]	Images	253 (images)	232 (images)	Senior clinician
Elsawy A et al. (2021) [32]	Images	29172 (train), 7293 (val)	23760	Certified cornea specialist
Khan ZK et al. (2021)[33]	Images	90	22	Clinician
Xiao P et al. (2021) [34]	Images	15	Nan	2 ophthalmologists
Yeh C-H et al. (2021) [35]	Images	398 (train), 99 (val)	209	Trained clinician
da Cruz LB et al. (2020)[36]	Tabular	106 (10-fold CV)	Nan	Optometrist
da Cruz LB et al. (2020) [37]	Tabular	106 (10-fold CV)	Nan	Optometrist
Fu P-I et al. (2020) [38]	Tabular	28	Nan	Nan (clinical test results, subjective report)
Fujimoto K et al. (2020) [39]	Tabular	195	Nan	Nan (kerato-conjunctival staining for DED)
Maruoka S et al. (2020) [40]	Images	221 (5-fold CV)	Nan	3 eyelid specialists
Prabhu SM et al.(2020) [41]	Images	600	200	Clinical experts
Stegmann H et al. (2020) [42]	Images	6658 (images) (5-fold CV)	Nan	Experienced investigator
Wei S et al. (2020) [43]	Images	5000*	53 ($3-5$ per patient)	Experienced investigator
Giannaccare G et al. (2019) [44]	Tabular	Nan	69	Experienced investigator [45]
Llorens-Quintana et al. (2020) [47]	Images	149	Nan	Clinicians
Wang J et al. (2019) [48]	Images	398 (train) 99 (val)	209	Experienced clinician
Yabusaki K et al. (2019) [49]	Tabular	93**	45**	Skilled ophthalmologist
Yang J et al. (2019) [50]	Images	520	Nan	ImageJ software
Szyperski PD (2018) [51]	Tabular	110	Nan	Nan
Hwang H et al. (2017) [52]	Frames	34	Nan	Meibomian gland expert
Koprowski R et al. (2017) [53]	Images	228 (images)	Nan	Specialized clinicians
Peteiro-Barral D et al. (2017) [54]	Tabular	105 (LOO CV)	Nan	Experts
Koprowski R et al. (2016) [55]	Images	172 (images)	Nan	Ophthalmology expert
Remeseiro B et al. (2016) [56]	Tabular	Nan	128	Optometrists
Remeseiro B et al. (2016) [57]	Tabular	Sampled from test set	50	4 optometrists
Kanellopoulos AJ et al. (2014) [58]	Tabular	140	Nan	Ophthalmologist
Ramos L et al. (2014) [59]	Videos	18	Nan	2/4 experts
Ramos L et al. (2014) [60]	Videos	12	6	4 experts
Remeseiro et al. (2014) [61]	Tabular	511 (10-fold CV)	Nan	Experts
García-Resúa C et al. (2013) [62]	Tabular	105 (6-fold CV)	Nan	Experienced investigator
Rodriguez R et al. (2013) [63]	Tabular	99 (images)	Nan	5 trained investigators
Koh YW et al. (2012) [64]	Tabular	28***	27***	Experts
Yedidya T et al. (2009) [65]	Videos	22	Nan	Clinician
Yedidya T et al. (2007) [13]	Frames	8****	Nan	Optometrist (evaluated 3 of the 8 patients)
Mathers WD et al. (2004) [66]	Tabular	513 (10-fold CV)	Nan	Nan (clinical test results)
Biochemical Investigations				
Cartes C et al. (2019) [67]	Tabular	40 (noise added)	40 (no noise)	Nan (clinical test results, subjective report)
Jung JH et al. (2017) [68]	Tabular	10	Nan	Ophthalmologist
Gonzalez N et al. (2014) [69]	Tabular	70% of 93**	30% of 93**	Nan (clinical tests)
Grus FH et al. (2005) [70]	Tabular	50 % of 159	50 % of 159	Nan (clinical test results, subjective report)
Grus FH et al. (1999) [71]	Tabular	30	30	Nan (clinical test results, subjective report)
Grus FH et al. (1998) [72]	Tabular	119	$\oplus$	Nan (clinical test results, subjective report)
Demographical Investigations				
Choi HR et al. (2020) [73]	Tabular	2272	Nan	Nan (subjective report)
Nam SM et al. (2020) [74]	Tabular	80 % of 4391	20 % of 4391	Ophtalmologist
Kaido M et al. (2015) [75]	Tabular	369	Nan	Dry eye specialists

Abbreviations: Nan = not available; val = validation; CV = crossvalidation; DED = dry eye disease; LOO = leave one out; * = pretraining images; ** = randomly selected samples, process repeated 10 times; *** = randomly selected samples, process repeated 100 times; **** = $3-5$ sequences of video per patient; $\oplus$ = For multivariate analysis model, but the number of samples was not mentioned

4.2. Fluorescein tear break-up time

Shorter break-up time indicates an unstable tear film and higher probability of DED. Machine learning has been employed to detect dry areas in TBUT videos and estimate TBUT [13, 65, 59, 60]. Use of the Levenberg-Marquardt algorithm to detect dry areas achieved an accuracy of 91% compared to assessments by an optometrist [13]. Application of Markov random fields to label pixels based on degree of dryness was used to estimate TBUT resulting in an average difference of 2.34 seconds compared to clinician assessments [65]. Polynomial functions have also been used to determine dry areas, where threshold values were fine-tuned before estimation of TBUT [59]. This method resulted in more than 90% of the videos deviating by less than ± 2.5 seconds compared to analyses done by four experts on videos not used for training [60]. Taken together, these studies indicate that TBUT values obtained using automatic methods are within an acceptable range compared to experts. However, we only found four studies, all of them including a small number of subjects. Further studies are needed to verify the findings and to test models on external data.

4.3. Interferometry and slit-lamp images

Interferometry is a useful tool that gives a snapshot of the status of the tear film lipid layer, which can be used to aid diagnosis of DED. Machine learning systems have been applied to interferometry and slit-lamp images for lipid layer classification based on morphological properties [62, 61, 57, 56, 54, 36, 37], estimation of the lipid layer thickness [52, 38], diagnosis of DED [51, 49], determination of ocular redness [63] and estimation of tear meniscus height [50, 31].

Diagnosis of DED can be based on the following morphological properties: open meshwork, closed meshwork, wave, amorphous and color fringe [76]. Most studies used these properties to automatically classify interferometer lipid layer images using machine learning. Garcia et al. used a K-nearest neighbors model trained to classify images resulting in an accuracy of 86.2% [62]. Remeseiro et al. explored various support vector machine (SVM) models for use in final classification [61, 57, 56]. In one of the studies, the same data was used for training and testing, which is not ideal [57]. Another study did not report the data their system was trained on [56]. Peteiro et al. evaluated images using five different machine learning models [54]. In this study, the amorphous property was not included as one of possible classifications, as opposed to the other studies. A simple neural network achieved the overall best performance with an accuracy of 96%. However, because leave-one-out cross validation was applied, the model may have overfitted on the training data [21]. da Cruz et al. compared six different machine learning models and found that the random forest was the best classifier, regardless of the pre-processing steps used [36, 37]. The highest performance was achieved by application of Ripley's K function in the image pre-processing phase, and Greedy Stepwise technique used simultaneously with the machine learning models for feature selection [37]. Since all models were evaluated

with cross validation, the system should be externally evaluated on new images before being considered for routine use in the clinic.

Hwang et al. investigated whether tear film lipid layer thickness can be used to distinguish meibomian gland dysfunction (MGD) severity groups [52]. Machine learning was used to estimate the thickness from Lipiscanner and slit-lamp videos with promising results. Images were pre-processed and the flood-fill algorithm and canny edge detection were applied to locate and extract the iris from the pupil. A significant difference between two MGD severity groups was detected, suggesting that the technique could be used for the evaluation of MGD. Keratograph images can also be used to determine tear film lipid layer thickness. Comparison of two different image analysis methods using a generalized linear model showed that there was a a high correlation between the two techniques [38]. The authors concluded that the simple technique was sufficient for evaluation of tear film lipid layer thickness. However, only 28 subjects were included in the study.

The use of fractal dimension estimation techniques was investigated for feature extraction from interferometer videos for diagnosis of DED [51]. The technique was found to be fast and had an area under the receiver operating characteristic curve (AUC) value of 0.786, compared to a value of 0.824 for an established method (See Appendix A.1 and Figure A.4a for a description of the receiver operating characteristic curve). Tear film lipid interferometer images were analysed using an SVM [49]. Extracted features from the images were passed to the SVM model, which classified the images as either healthy, aqueous-deficient DED, or evaporative DED. The agreement between the model and a trained ophthalmologist was high, with a reported Kappa value of 0.82. The model performed best when detecting aqueous-deficient DED.

Ocular redness is an important indicator of dry eyes. Only one of the reviewed studies described an automated system for evaluation of ocular redness associated with DED [63]. Slit-lamp images were acquired from 26 subjects with a history of DED. Features representing the ocular redness intensity and horizontal vascular component were extracted with a Sobel operator. A multiple linear regression model was trained to predict ocular redness based on the extracted features. The system achieved an accuracy of 100%. The authors suggested that an objective system like this could replace subjective gradings by clinicians in multicentered clinical studies.

The tear meniscus contains $75 - 90\%$ of the aqueous tear volume [77]. Consequently, the tear meniscus height can be used as a quantitative indicator for DED caused by aqueous deficiency. When connected component labelling was applied to slit-lamp images, the Pearson's correlation between the predicted meniscus heights and an established software methodology (ImageJ [78]) was high, ranging between 0.626 and 0.847 [50]. The machine learning system was found to be more accurate than four experienced ophthalmologists. The tear meniscus height can also be estimated from keratography images using a CNN [31]. The automatic

machine learning system achieved an accuracy of 82.5% and was found to be more effective and consistent than a well-trained clinician working with limited time.

Many of the studies apply SVM as their type of machine learning model without testing how other machine learning models perform. However, three of the studies tested several types of models and found that SVM did not perform the best [54, 36, 37]. It is difficult to compare the studies due to different applications and evaluation metrics. Despite promising results, most of the studies [62, 61, 57, 54, 36, 37, 52, 38, 51, 63, 50] did not evaluate their systems on external data. The systems should be tested on independent data before they can be considered for clinical application. Moreover, some studies were small [63, 38] or pilots [50, 31], and the suggested models should be tested on a larger number of subjects.

4.4. In vivo confocal microscopy

IVCM is a valuable non-invasive tool used to examine the corneal nerves and other features of the cornea [79]. IVCM images were used in a small study to assess characteristics of the corneal subbasal nerve plexus for diagnosis of DED [44]. Application of random forest and a deep neural network [45] gave promising results with an AUC value of 0.828 for detecting DED [44]. IVCM images of corneal nerves can also be analyzed by machine learning models to estimate the length of the nerve fiber [43]. Authors used a CNN with a U-net architecture that had been pre-trained on more than 5,000 IVCM images of corneal nerves. The model showed that nerve fiber length was significantly longer after intense pulsed light treatment in MGD patients, which agreed with manual annotations from an experienced investigator with an AUC value of 0.96 and a sensitivity of 0.96. High-resolution IVCM images were also used to detect obstructive MGD [40]. Combinations of nine different CNNs were trained and tested on the images using 5-fold cross validation. Classification by the models was compared to diagnosis made by three eyelid specialists. The best performance was achieved when four different models were combined, with high sensitivity, specificity and AUC values, see Table 1. These promising results suggest that CNNs can be useful for detection and evaluation of MGD. Deep learning methods such as CNNs have the advantage that feature extraction from the images prior to analysis is not required as this is performed automatically by the model.

IVCM images have been investigated for changes in immune cells across different severities of DED for diagnostic purposes [30]. A generalized linear model showed significant differences in dendritic cell density and morphology between DED patients and healthy individuals, but not between the different DED subgroups, see Table 1. While results using machine learning to interpret IVCM images are promising, larger clinical studies are needed to validate findings before clinical use can be considered.

4.5. Meibography

The meibomian glands are responsible for producing meibum, important for protecting the tear fluid from evaporation. Reduced secretion of meibum due to a reduced number of functional meibomian glands and/or obstruction of the ducts is a major cause of evaporative DED and MGD. Meibography is a common technique for diagnosing MGD [80]. Classification of meibomian glands using meibography is routine for experienced experts, but this is not the case for all clinicians. Moreover, automatic methods can be faster than human assessment.

Meibography images may require several pre-processing steps before they can be classified. One study trained an SVM on extracted features from the images [64]. Pre-processing included the dilation, flood-fill, skeletonization and pruning algorithms. The model achieved a sensitivity of 0.979 and specificity of 0.961. However, in contrast to all other image analysis methods, this method is not completely automatic as the images need to be manipulated manually before they are passed on to the system.

A combination of Otsu's method and the skeletonization and watershed algorithms was useful in automatically quantifying meibomian glands [55]. This method was faster than an ophthalmologist and achieved a sensitivity and specificity of 0.993 and 0.975, respectively. Another automatic method applied Bézier curve fitting as part of the analysis [53]. The reported sensitivity was 1.0, while the specificity was 0.98. Xiao et al. sequentially applied a Prewitt operator, Graham scan, fragmentation and skeletonization algorithms for image analysis to quantify meibomian glands [34]. The agreement between the model results and two ophthalmologists was high with Kappa values larger than 0.8 and low false positive rates (< 0.06). The false negative rate was 0.19, suggesting that some glands were missed by the method. A considerable weakness of this study was that only 15 images were used for model development, and consequently it might not work well on unseen data. Another study automatically graded MGD severity using a Sobel operator, polynomial functions, fragmentation algorithm and Otsu's method [47]. While the method was found to be faster, the results were significantly different from clinician assessments.

Deep learning approaches were used by four studies evaluating meibomian gland features [48, 41, 35, 33]. These systems are fully automated and apply some of the latest technologies within image analysis. Wang et al. used four different CNNs to determine meibomian gland atrophy [48]. The CNNs were trained to identify meibomian gland drop-out areas and estimate the percentage atrophy in a set of images. Comparison of model predictions with experienced clinicians indicated that the best CNN (ResNet50 architecture) was superior. Yeh et al. developed a method to evaluate meibomian gland atrophy by extracting features from meibography images with a special type of unsupervised CNN before application of a K-nearest neighbors model to allocate a meiboscore [35]. The system achieved an accuracy of 80.9%, outperforming annotations by the clinical team. Moreover, hierarchical clustering of the extracted features from the CNN could show

relationships between meibography images. Another study used a CNN to automatically assess meibomian gland characteristics [41]. Images from two different devices collected from various hospitals were used to train and evaluate the CNN. This is an example of uncommonly good practice, as most medical AI systems are developed and evaluated on data from only one device and/or hospital. The only study to use a GAN architecture tested it on infrared 3D images of meibomian glands in order to evaluate MGD [33]. Comparing the model output with true labels, the performance scores were better than for state of the art segmentation methods. The Pearson correlations between the new automated method and two clinicians were 0.962 and 0.968.

Four of the studies did not evaluate their proposed systems on external data [55, 53, 47, 34]. Since the number of images used for model development was limited, the models can have overfit, and external evaluations should be performed to test how well the systems generalize to new data.

4.6. Tear osmolarity

Tear osmolarity is a measure of tear concentration, and high values can indicate dry eyes. Cartes et al. [67] investigated use of machine learning to detect DED based on this test. Four different machine learning models were compared. Noise was added to osmolarity measurements during the training phase, while original data without noise was used for final evaluation. The logistic regression model achieved 85% accuracy. However, since the models were trained and tested on the same data, the reported score is most likely not representative for how well the model generalizes to new data.

4.7. Proteomic analysis

Proteomic analysis describes the qualitative and quantitative composition of proteins present in a sample. Grus et al. compared tear proteins in individuals with diabetic DED, non-diabetic DED and healthy controls for discrimination between the groups [72]. The authors used discriminant analysis and principal component analysis combined with k-means clustering. Both models achieved low accuracies when predicting all three categories. However, classification into DED and non-DED achieved accuracies of 72% and 71% for discriminant analysis and k-means clustering, respectively. In another study by the same group, tear proteins analyzed using deep learning discriminated subjects as healthy or having DED with an accuracy of 89% [71]. An accuracy of 71% was achieved using discriminant analysis. A combination of discriminant analysis for detecting the most important proteins and a deep neural network for classification was also investigated [70]. High accuracy, sensitivity and specificity were reported. Discriminant analysis was also used by Gonzalez et al. in analysis of the tear proteome [69]. The most important proteins were selected to train an artificial neural network to classify tear samples as aqueous-deficient DED, MGD or healthy. The model gave an overall accuracy of 89.3%. Principal component analysis yielded good separation of healthy controls, aqueous-deficient

DED and MGD data-points, indicating that the proteins were good candidates for classification of the three conditions. This system achieved the highest accuracy of all the reviewed proteomic studies. Considered together, the results from the four studies [72, 71, 70, 69] suggest that neural networks applied alone or together with other techniques perform better than discriminant analysis for detecting DED-related protein patterns in the tear proteome.

Jung et al. used a network model based on modularity analysis to describe the tear proteome with respect to immunological and inflammatory responses related to DED [68]. In this study, patterns in tears and lacrimal fluid were investigated in patients with DED. Since only 10 subjects were included, the study should be performed on a larger cohort of patients to verify the results.

4.8. Optical coherence tomography

Thickening of the corneal epithelium can be a sign of abnormalities in the cornea. Moreover, corneal thickness could potentially be a marker for DED. Kanellopoulos et al. developed a linear regression model to look for possible correlations between corneal thickness metrics measured using anterior segment optical coherence tomography (AS-OCT) and DED [58]. However, neither the model predictions nor performance were reported, making it difficult to assess the usefulness of the study. The type of instrument used to determine the corneal thickness was found to affect the results [39]. Measurements from AS-OCT and Pentacam were compared and multivariable regression was used to detect differences between the two techniques regarding the measured central corneal thickness and the thinnest corneal thickness. Individuals with mild DED, severe DED and healthy subjects were examined. The two techniques gave significantly different results in terms of the resulting β-coefficients in the multivariable regression model for individuals with severe DED. Images from clinical examinations with AS-OCT were used to diagnose DED [32]. A pretrained VGG19 CNN [81] was fine-tuned using separate images for training and validation. Two similar CNN models were developed, and evaluation was performed on an external test set. Both achieved impressively high performance scores. The AUC values were 0.99 and 0.98. This is one out of two studies in this review that used an independent test sets after model development. Such practice is essential for a realistic impression of how well the model generalizes to new data not used during model development. The good performance is likely linked to the large amounts of training data ($29,000$ images), which is essential for deep learning methods. Most of the reviewed studies use significantly smaller data sets, which constitutes a disadvantage. Stegmann et al. analysed OCT images from healthy subjects for automatic detection of the lower tear meniscus [42]. Two different CNNs were trained and evaluated using 5-fold cross validation. The tear menisci detected by the models were compared to evaluations from an experienced grader. The best CNN achieved an average accuracy of 99.95%, sensitivity of 0.9636 and specificity of 0.9998. The system is promising regarding fast and accurate

segmentation of OCT images. However, more images from different OCT systems, including non-healthy subjects, should be used to verify and improve the analysis.

The two studies [81, 42] showed that CNNs could be an appropriate tool for image analysis. CNNs are likely to increase in popularity within the field of DED due to promising results for solving image related tasks, including feature extraction.

4.9. Other clinical tests

Machine learning models were used to analyse results from a variety of clinical tests to expand understanding of the DED process [66]. The study included subjects with DED and healthy subjects. Subjective cutoff values from clinical tests were used to assign subjects to the DED class. Hierarchical clustering and a decision tree were applied sequentially to group the subjects based on their clinical test results. The resulting groups were compared to the original groups. Because the analysis was based on objective measurements, it could be used to develop more objective diagnostic criteria. This could lead to earlier detection and more effective treatment of DED.

4.10. Population surveys

Population surveys can provide valuable insight regarding the prevalence of DED and help detect risk factors for developing the disease. Japanese visual terminal display workers were surveyed with the objective of detecting DED [75]. Dry eye exam data and subjective reports were used for diagnosis. This was passed to a discriminant analysis model. When compared to diagnosis by a dry eye specialist, the model showed a high sensitivity of 0.931, but low specificity of 0.437. This is a very low specificity, but is not necessarily bad if the aim is to detect as many cases of DED as possible and there is less concern about misclassification of healthy individuals. Data from a national health survey were analysed in order to detect risk factors for DED [74]. Here, individuals were regarded as having DED if they had been diagnosed by an ophthalmologist, and were experiencing dryness. Feature modifications were performed by a decision tree, and the most important features were selected using lasso. β-coefficients from a logistic regression trained on the most important features were used to rank the features. Women, individuals who had received refractive surgery and those with depression were detected as having the highest risk for developing DED. Even though the models in the study were trained on data from more than 3500 participants, the reported performance scores were among the poorest in this review with a sensitivity of 0.66 and a specificity of 0.68. A possible reason could be that the selected features were not ideal for detecting DED. However, the detected risk factors have previously been shown to be associated with DED [3, 82, 83]. The findings suggest that the data quality from population surveys might not be as high as in other types of studies, which could lead to misinterpretation by the machine learning model.

The association between DED and dyslipidemia was investigated by combining data from two population surveys in Korea in [73]. A generalized linear model was used to investigate linear characteristics between features and the severity of DED. The model showed significant increase in age, blood pressure and prevalence of hypercholesterolemia over the range from no DED to severe DED. Evaluation of the association between dyslipidemia and DED using linear regression showed that the odds ratio for men with dyslipidemia was higher than 1 compared to men without dyslipidemia. This association was not found in women. The study results suggest a positive association between DED and dyslipidemia in men, but not in women.

4.11. Future perspectives

In order to benchmark existing and future models, we advocate that the field of DED should have a common, centralized and openly available data set for testing and evaluation. The data should be fully representative for the relevant clinical tests. In order to ensure that models are applicable to all populations of patients, medical institutions, and types of equipment around the world, they must be evaluated on data from different demographic groups of patients across several clinics and, if relevant, from different medical devices. Moreover, the test data set should not be available for model development, but only for final evaluation. A common standard on these processes will increase the reproducibility and comparability of studies. Standardized collection and handling of clinical data and samples would also facilitate comparisons between different instruments and clinics [84]. In addition, a cross hospitals/centers data set would solve important challenges of applying AI in clinical practice, such as metrics not reflecting clinical applicability, difficulties in comparing algorithms, and underspecification. These have all been identified as being among the main obstacles for adoption of any medical AI system in clinical practice [85, 86].

A possible challenge regarding implementation in the clinic is that hospitals do not necessarily use the same data platforms, which might prevent widespread use of machine learning systems. Consequently, solutions for implementing digital applications across hospitals should be considered.

Model explanations are important in order to understand why a complex machine learning model produces a certain prediction. For healthcare providers to trust the systems and decide to use them in the clinic, the systems should provide understandable and sound explanations of the decision-making process. Moreover, they could assist clinicians when making medical decisions [18]. When developing new machine learning systems within DED, effort should be made to present the workings of the resulting models and their predictions in an easy to interpret fashion.

5. Conclusions

We observed a large variation in the type of clinical tests and the type of data used in the reviewed studies. This is also true regarding the extent of pre-processing applied to the data before passing it to the machine learning models. The studies analysing images can be divided into those applying deep learning techniques directly on the images, and those performing extensive pre-processing and feature extraction before the data is passed to the machine learning model in a tabular format. The number of studies belonging to the first group has increased significantly over the past 3 years. As deep learning techniques become more established, these will probably replace more traditional image pre-processing and feature extraction techniques.

We noted that there was a lack of consensus regarding how best to perform model development, including evaluation. This made it difficult to estimate how well some models will perform in the clinic and with new patients, and also to compare the different models. Comparison was further complicated by the use of different types of performance scores. In addition there was no culture of data and code sharing, which makes reproducibility of the results impossible. For the future, focus should be put on establishing data and code sharing as a standard procedure.

In conclusion, the results from the different studies' machine learning models are promising, although much work is still needed on model development, clinical testing and standardisation. AI has a high potential for use in many different applications related to DED, including automatic detection and classification of DED, investigation of the etiology and risk factors for DED, and in the detection of potential biomarkers. Effort should be made to create common guidelines for the model development process, especially regarding model evaluation. Prospective testing is recommended in order to evaluate whether proposed models can improve the diagnostics of DED, and the health and quality of life of patients with DED.

Disclosure

The authors report no conflicts of interest.

References

[1] F. Stapleton, M. Alves, V. Y. Bunya, I. Jalbert, K. Lekhanont, F. Malet, K.-S. Na, D. Schaumberg, M. Uchino, J. Vehof, et al., TFOS DEWS II epidemiology report, The ocular surface 15 (3) (2017) 334–365. doi:http://dx.doi.org/10.1016/j.jtos.2017.05.003.

[2] G. Geerling, J. Tauber, C. Baudouin, E. Goto, Y. Matsumoto, T. O'Brien, M. Rolando, K. Tsubota, K. K. Nichols, The international workshop on meibomian gland dysfunction: report of the subcommittee

on management and treatment of meibomian gland dysfunction, Investigative ophthalmology & visual science 52 (4) (2011) 2050–2064. doi:http://dx.doi.org/10.1167/iovs.10-6997g.

[3] C. Matossian, M. McDonald, K. E. Donaldson, K. K. Nichols, S. MacIver, P. K. Gupta, Dry eye disease: consideration for women's health, Journal of Women's Health 28 (4) (2019) 502–514. doi:http://dx.doi.org/10.1089/jwh.2018.7041.

[4] J. J. Nichols, C. Ziegler, G. L. Mitchell, K. K. Nichols, Self-reported dry eye disease across refractive modalities, Investigative ophthalmology & visual science 46 (6) (2005) 1911–1914. doi:http://dx.doi.org/10.1167/iovs.04-1294.

[5] X. Zhang, L. Zhao, S. Deng, X. Sun, N. Wang, Dry eye syndrome in patients with diabetes mellitus: prevalence, etiology, and clinical characteristics, Journal of ophthalmology 2016 (2016). doi:http://dx.doi.org/10.1155/2016/8201053.

[6] J. T. Mandell, M. Idarraga, N. Kumar, A. Galor, Impact of air pollution and weather on dry eye, Journal of Clinical Medicine 9 (11) (2020). doi:http://dx.doi.org/10.3390/jcm9113740.
URL https://www.mdpi.com/2077-0383/9/11/3740

[7] M. D. Willcox, P. Argüeso, G. A. Georgiev, J. M. Holopainen, G. W. Laurie, T. J. Millar, E. B. Papas, J. P. Rolland, T. A. Schmidt, U. Stahl, et al., TFOS DEWS II tear film report, The ocular surface 15 (3) (2017) 366–403. doi:http://dx.doi.org/10.1016/j.jtos.2017.03.006.

[8] J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, A proposal for the Dartmouth summer research project on Artificial Intelligence, august 31, 1955, AI magazine 27 (4) (2006) 12–12. doi:http://dx.doi.org/10.1609/aimag.v27i4.1904.

[9] S. Legg, M. Hutter, Universal intelligence: A definition of machine intelligence, Minds and machines 17 (4) (2007) 391–444. doi:http://dx.doi.org/10.1007/s11023-007-9079-x.

[10] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, H. Bogunović, Artificial Intelligence in retina, Progress in retinal and eye research 67 (2018) 1–29. doi:http://dx.doi.org/10.1016/j.preteyeres.2018.07.004.

[11] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature medicine 24 (9) (2018) 1342–1350. doi:http://dx.doi.org/10.1038/s41591-018-0107-6.

[12] K. Ciężar, M. Pochylski, 2D fourier transform for global analysis and classification of meibomian gland images, The Ocular Surface 18 (4) (2020) 865–870. doi:https://doi.org/10.1016/j.jtos.2020.09.005.
URL https://www.sciencedirect.com/science/article/pii/S1542012420301452

[13] T. Yedidya, R. Hartley, J.-P. Guillon, Y. Kanagasingam, Automatic dry eye detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2007, pp. 792–799. doi:http://dx.doi.org/10.1007/978-3-540-75757-3_96.

[14] K. B. Nielsen, M. L. Lautrup, J. K. Andersen, T. R. Savarimuthu, J. Grauslund, Deep learning–based algorithms in screening of diabetic retinopathy: A systematic review of diagnostic performance, Ophthalmology Retina 3 (4) (2019) 294–304. doi:http://dx.doi.org/10.1016/j.oret.2018.10.014.

[15] E. Pead, R. Megaw, J. Cameron, A. Fleming, B. Dhillon, E. Trucco, T. MacGillivray, Automated detection of age-related macular degeneration in color fundus photography: a systematic review, Survey of Ophthalmology 64 (4) (2019) 498–511. doi:10.1016/j.survophthal.2019.02.003.
URL https://www.sciencedirect.com/science/article/pii/S0039625718302078

[16] R. H. Gensure, M. F. Chiang, J. P. Campbell, Artificial Intelligence for retinopathy of prematurity, Current Opinion in Ophthalmology 31 (5) (2020) 312–317. doi:http://dx.doi.org/10.1097/ICU.0000000000000680.

[17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature (London) 542 (2017) 115–118. doi:http://dx.doi.org/10.1038/nature21056.

[18] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature biomedical engineering 2 (2018) 749–760. doi:https://doi.org/10.1038/s41551-018-0304-0.

[19] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402–2410. doi:http://dx.doi.org/10.1001/jama.2016.17216.

[20] S. Yousefi, H. Takahashi, T. Hayashi, H. Tampo, S. Inoda, Y. Arai, H. Tabuchi, P. Asbell, Predicting the likelihood of need for future keratoplasty intervention using artificial intelligence, The Ocular Surface

18 (2) (2020) 320–325. doi:https://doi.org/10.1016/j.jtos.2020.02.008.
URL https://www.sciencedirect.com/science/article/pii/S1542012420300276

[21] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009. doi:https://doi.org/10.1111/j.1467-985X.2010.00646_6.x.

[22] J.-O. Palacio-Niño, F. Berzal, Evaluation metrics for unsupervised learning algorithms (2019). arXiv:1905.05667.

[23] C. Le Berre, W. J. Sandborn, S. Aridhi, M.-D. Devignes, L. Fournier, M. Smaïl-Tabbone, S. Danese, L. Peyrin-Biroulet, Application of artificial intelligence to gastroenterology and hepatology, Gastroenterology 158 (1) (2020) 76–94. doi:https://doi.org/10.1053/j.gastro.2019.08.058.

[24] J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, J. Brink, Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success, Journal of the American College of Radiology 15 (3) (2018) 504–508. doi:https://doi.org/10.1016/j.jacr.2017.12.026.

[25] V. L. Thambawita, I. Strümke, S. Hicks, M. A. Riegler, P. Halvorsen, S. Parasa, Data augmentation using generative adversarial networks for creating realistic artificial colon polyp images: Validation study by endoscopists, Gastrointestinal Endoscopy 93 (6) (2021) AB190.

[26] M. A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, Potential biases in machine learning algorithms using electronic health record data, JAMA internal medicine 178 (11) (2018) 1544–1547. doi:https://10.1001/jamainternmed.2018.3763.

[27] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, O'Reilly Media, 2019.

[28] European Commission, Proposal for a regulation laying down harmonised rules on Artificial Intelligence (4 2021).
URL https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[29] U.S. Food & Drug Administration, Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device (SaMD) Action Plan (1 2021).
URL https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[30] S. Aggarwal, A. Kheirkhah, B. M. Cavalcanti, A. Cruzat, A. Jamali, P. Hamrah, Correlation of corneal immune cell changes with clinical severity in dry eye disease: An in vivo confocal microscopy study, The Ocular Surface 19 (2021) 183–189. doi:http://dx.doi.org/10.1016/j.jtos.2020.05.012.
URL https://www.sciencedirect.com/science/article/pii/S1542012420300963

[31] X. Deng, L. Tian, Z. Liu, Y. Zhou, Y. Jie, A deep learning approach for the quantification of lower tear meniscus height, Biomedical Signal Processing and Control 68 (2021) 102655. doi:https://doi.org/10.1016/j.bspc.2021.102655.
URL https://www.sciencedirect.com/science/article/pii/S1746809421002524

[32] A. Elsawy, T. Eleiwa, C. Chase, E. Ozcan, M. Tolba, W. Feuer, M. Abdel-Mottaleb, M. Abou Shousha, Multidisease deep learning neural network for the diagnosis of corneal diseases, American Journal of Ophthalmology 226 (2021) 252–261. doi:https://doi.org/10.1016/j.ajo.2021.01.018.
URL https://www.sciencedirect.com/science/article/pii/S0002939421000398

[33] Z. K. Khan, A. I. Umar, S. H. Shirazi, A. Rasheed, A. Qadir, S. Gul, Image based analysis of meibomian gland dysfunction using conditional generative adversarial neural network, BMJ Open Ophthalmology 6 (1) (2021). arXiv:https://bmjophth.bmj.com/content/6/1/e000436.full.pdf, doi:http://dx.doi.org/10.1136/bmjophth-2020-000436.
URL https://bmjophth.bmj.com/content/6/1/e000436

[34] P. Xiao, Z. Luo, Y. Deng, G. Wang, J. Yuan, An automated and multiparametric algorithm for objective analysis of meibography images, Quantitative Imaging in Medicine and Surgery 11 (4) (2021) 1586–1599. doi:http://dx.doi.org/10.21037/qims-20-611.

[35] C.-H. Yeh, S. X. Yu, M. C. Lin, Meibography phenotyping and classification from unsupervised discriminative feature learning, Translational Vision Science & Technology 10 (2) (2021) 4–4. arXiv:https://arvojournals.org/arvo/content_public/journal/tvst/938516/i2164-2591-10-2-4_1612519083.80616.pdf, doi:https://doi.org/10.1167/tvst.10.2.4.

[36] L. B. da Cruz, J. C. Souza, J. A. de Sousa, A. M. Santos, A. C. de Paiva, J. D. S. de Almeida, A. C. Silva, G. B. Junior, M. Gattass, Interferometer eye image classification for dry eye categorization using phylogenetic diversity indexes for texture analysis, Computer Methods and Programs in Biomedicine 188 (2020) 105269. doi:https://doi.org/10.1016/j.cmpb.2019.105269.
URL https://www.sciencedirect.com/science/article/pii/S0169260719310995

[37] L. B. da Cruz, J. C. Souza, A. C. de Paiva, J. D. S. de Almeida, G. B. Junior, K. R. T. Aires, A. C. Silva, M. Gattass, Tear film classification in interferometry eye images using phylogenetic diversity

indexes and ripley's k function, IEEE Journal of Biomedical and Health Informatics 24 (12) (2020) 3491–3498. doi:http://dx.doi.org/10.1109/JBHI.2020.3026940.

[38] P.-I. Fu, P.-C. Fang, R.-W. Ho, T.-L. Chao, W.-H. Cho, H.-Y. Lai, Y.-T. Hsiao, M.-T. Kuo, Determination of tear lipid film thickness based on a reflected placido disk tear film analyzer, Diagnostics 10 (6) (2020). doi:https://doi.org/10.3390/diagnostics10060353.
URL https://www.mdpi.com/2075-4418/10/6/353

[39] K. Fujimoto, T. Inomata, Y. Okumura, N. Iwata, K. Fujio, A. Eguchi, K. Nagino, H. Shokirova, M. Karasawa, A. Murakami, Comparison of corneal thickness in patients with dry eye disease using the pentacam rotating scheimpflug camera and anterior segment optical coherence tomography, PLOS ONE 15 (2) (2020) e0228567. doi:http://dx.doi.org/10.1371/journal.pone.0228567.

[40] S. Maruoka, H. Tabuchi, D. Nagasato, H. Masumoto, T. Chikama, A. Kawai, N. Oishi, T. Maruyama, Y. Kato, T. Hayashi, C. Katakami, Deep neural network-based method for detecting obstructive meibomian gland dysfunction with in vivo laser confocal microscopy, Cornea 39 (6) (2020) 720–725. doi:https://doi.org/10.1097/ICO.0000000000002279.

[41] S. M. Prabhu, A. Chakiat, S. S, K. P. Vunnava, R. Shetty, Deep learning segmentation and quantification of meibomian glands, Biomedical signal processing and control 57 (2020) 101776. doi:https://doi.org/10.1016/j.bspc.2019.101776.
URL https://www.sciencedirect.com/science/article/pii/S174680941930357X

[42] H. Stegmann, R. M. Werkmeister, M. Pfister, G. Garhöfer, L. Schmetterer, V. A. dos Santos, Deep learning segmentation for optical coherence tomography measurements of the lower tear meniscus, Biomedical Optics Express 11 (3) (2020) 1539–1554. doi:https://doi.org/10.1364/BOE.386228.
URL http://www.osapublishing.org/boe/abstract.cfm?URI=boe-11-3-1539

[43] S. Wei, X. Ren, Y. Wang, Y. Chou, X. Li, Therapeutic effect of intense pulsed light (ipl) combined with meibomian gland expression (mgx) on meibomian gland dysfunction (mgd), Journal of ophthalmology 2020 (2020). doi:http://dx.doi.org/10.1155/2020/3684963.

[44] G. Giannaccare, M. Pellegrini, S. Sebastiani, F. Moscardelli, P. Versura, E. C. Campos, In vivo confocal microscopy morphometric analysis of corneal subbasal nerve plexus in dry eye disease using newly developed fully automated system, Graefe's archive for clinical and experimental ophthalmology 257 (3) (2019) 583–589. doi:https://doi-org.ezproxy.uio.no/10.1007/s00417-018-04225-7.

[45] X. Chen, J. Graham, M. A. Dabbah, I. N. Petropoulos, M. Tavakoli, R. A. Malik, An automatic tool for quantification of nerve fibers in corneal confocal microscopy images, IEEE Transactions on Biomedical Engineering 64 (4) (2017) 786–794. doi:https://doi.org/10.1109/TBME.2016.2573642.

[46] M. Dabbah, J. Graham, I. Petropoulos, M. Tavakoli, R. Malik, Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging, Medical Image Analysis 15 (5) (2011) 738–747. doi:https://doi.org/10.1016/j.media.2011.05.016.
URL https://www.sciencedirect.com/science/article/pii/S1361841511000806

[47] C. Llorens-Quintana, L. Rico-Del-Viejo, P. Syga, D. Madrid-Costa, D. R. Iskander, A novel automated approach for infrared-based assessment of meibomian gland morphology, Translational vision science & technology 8 (4) (2019) 17–17. doi:https://doi.org/10.1167/tvst.8.4.17.

[48] J. Wang, T. N. Yeh, R. Chakraborty, S. X. Yu, M. C. Lin, A deep learning approach for meibomian gland atrophy evaluation in meibography images, Translational Vision Science & Technology 8 (6) (2019) 37–37. doi:https://doi.org/10.1167/tvst.8.6.37.

[49] K. Yabusaki, R. Arita, T. Yamauchi, Automated classification of dry eye type analyzing interference fringe color images of tear film using machine learning techniques, Modeling and Artificial Intelligence in Ophthalmology 2 (3) (2019) 28–35. doi:https://doi.org/10.35119/maio.v2i3.90.

[50] J. Yang, X. Zhu, Y. Liu, X. Jiang, J. Fu, X. Ren, K. Li, W. Qiu, X. Li, J. Yao, TMIS: a new image-based software application for the measurement of tear meniscus height, Acta Ophthalmologica 97 (7) (2019) e973–e980. doi:http://dx.doi.org/10.1111/aos.14107.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/aos.14107

[51] P. D. Szyperski, Comparative study on fractal analysis of interferometry images with application to tear film surface quality assessment, Applied optics 57 (16) (2018) 4491–4498. doi:https://doi.org/10.1364/AO.57.004491.

[52] H. Hwang, H.-J. Jeon, K. C. Yow, H. S. Hwang, E. Chung, Image-based quantitative analysis of tear film lipid layer thickness for meibomian gland evaluation, Biomedical engineering online 16 (1) (2017) 1–15. doi:http://dx.doi.org/10.1186/s12938-017-0426-8.

[53] R. Koprowski, L. Tian, P. Olczyk, A clinical utility assessment of the automatic measurement method of the quality of meibomian glands, Biomedical engineering online 16 (82) (2017) 1–13. doi:https://doi.org/10.1186/s12938-017-0373-4.

[54] D. Peteiro-Barral, B. Remeseiro, R. Méndez, M. G. Penedo, Evaluation of an automatic dry eye test using mcdm methods and rank correlation, Medical & biological engineering & computing 55 (4) (2017) 527–536. doi:http://dx.doi.org/10.1007/s11517-016-1534-5.

[55] R. Koprowski, S. Wilczyński, P. Olczyk, A. Nowińska, B. Węglarz, E. Wylęgała, A quantitative method for assessing the quality of meibomian glands, Computers in Biology and Medicine 75 (2016) 130–138. doi:https://doi.org/10.1016/j.compbiomed.2016.06.001.
URL https://www.sciencedirect.com/science/article/pii/S0010482516301391

[56] B. Remeseiro, N. Barreira, C. García-Resúa, M. Lira, M. J. Giráldez, E. Yebra-Pimentel, M. G. Penedo, ideas: A web-based system for dry eye assessment, Computer Methods and Programs in Biomedicine 130 (2016) 186–197. doi:http://dx.doi.org/10.1016/j.cmpb.2016.02.015.
URL https://www.sciencedirect.com/science/article/pii/S0169260715301644

[57] B. Remeseiro, A. Mosquera, M. G. Penedo, CASDES: A computer-aided system to support dry eye diagnosis based on tear film maps, IEEE Journal of Biomedical and Health Informatics 20 (3) (2016) 936–943. doi:http://dx.doi.org/10.1109/JBHI.2015.2419316.

[58] A. J. Kanellopoulos, G. Asimellis, In vivo 3-dimensional corneal epithelial thickness mapping as an indicator of dry eye: Preliminary clinical assessment, American Journal of Ophthalmology 157 (1) (2014) 63–68.e2. doi:http://dx.doi.org/10.1016/j.ajo.2013.08.025.
URL https://www.sciencedirect.com/science/article/pii/S0002939413005850

[59] L. Ramos, N. Barreira, H. Pena-Verdeal, M. Giráldez, Automatic assessment of tear film break-up dynamics, Studies in health technology and informatics 207 (2014) 173–182. doi:http://dx.doi.org/10.3233/978-1-61499-474-9-173.

[60] L. Ramos, N. Barreira, A. Mosquera, M. Penedo, E. Yebra-Pimentel, C. García-Resúa, Analysis of parameters for the automatic computation of the tear film break-up time test based on cclru standards, Computer Methods and Programs in Biomedicine 113 (3) (2014) 715–724. doi:http://dx.doi.org/10.1016/j.cmpb.2013.12.003.
URL https://www.sciencedirect.com/science/article/pii/S0169260713003921

[61] B. Remeseiro, V. Bolon-Canedo, D. Peteiro-Barral, A. Alonso-Betanzos, B. Guijarro-Berdiñas, A. Mosquera, M. G. Penedo, N. Sánchez-Maroño, A methodology for improving tear film lipid layer classification, IEEE Journal of Biomedical and Health Informatics 18 (4) (2014) 1485–1493. doi:https://doi.org/10.1109/JBHI.2013.2294732.

[62] C. García-Resúa, M. J. G. Fernández, M. F. G. Penedo, D. Calvo, M. Penas, E. Yebra-Pimentel, New software application for clarifying tear film lipid layer patterns, Cornea 32 (4) (2013) 538–546. doi:http://dx.doi.org/10.1097/ICO.0b013e31824d0d04.

[63] J. D. Rodriguez, P. R. Johnston, G. W. Ousler, L. M. Smith, M. B. Abelson, Automated grading system for evaluation of ocular redness associated with dry eye, Clinical Ophthalmology 7 (2013) 1197 – 1204. doi:https://doi.org/10.2147/OPTH.S39703.

[64] Y. W. Koh, T. Celik, H. K. Lee, A. Petznick, L. H. Tong, Detection of meibomian glands and classification of meibography images, Journal of biomedical optics 17 (8) (2012) 086008. doi:https://doi.org/10.1117/1.JBO.17.8.086008.

[65] T. Yedidya, P. Carr, R. Hartley, J.-P. Guillon, Enforcing monotonic temporal evolution in dry eye images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2009, pp. 976–984. doi:https://doi.org/10.1007/978-3-642-04271-3_118.

[66] W. D. Mathers, D. Choi, Cluster analysis of patients with ocular surface disease, blepharitis, and dry eye, Archives of Ophthalmology 122 (11) (2004) 1700–1704. doi:http://dx.doi.org/10.1001/archopht.122.11.1700.
URL https://jamanetwork.com/journals/jamaophthalmology/articlepdf/416676/eeb30021.pdf

[67] C. Cartes, D. López, D. Salinas, C. Segovia, C. Ahumada, N. Pérez, F. Valenzuela, N. Lanza, R. L. Solís, V. Perez, et al., Dry eye is matched by increased intrasubject variability in tear osmolarity as confirmed by machine learning approach, Archivos de la Sociedad Española de Oftalmología (English Edition) 94 (7) (2019) 337–342. doi:http://dx.doi.org/10.1016/j.oftal.2019.03.007.

[68] J. H. Jung, Y. W. Ji, H. S. Hwang, J. W. Oh, H. C. Kim, H. K. Lee, K. P. Kim, Proteomic analysis of human lacrimal and tear fluid in dry eye disease, Scientific reports 7 (1) (2017) 1–11. doi:http://dx.doi.org/10.1038/s41598-017-13817-y.

[69] N. González, I. Iloro, J. Soria, J. A. Duran, A. Santamaría, F. Elortza, T. Suárez, Human tear peptide/protein profiling study of ocular surface diseases by spe-maldi-tof mass spectrometry analyses, EuPA Open Proteomics 3 (2014) 206–215. doi:https://doi.org/10.1016/j.euprot.2014.02.016.
URL https://www.sciencedirect.com/science/article/pii/S221296851400021X

[70] F. H. Grus, V. N. Podust, K. Bruns, K. Lackner, S. Fu, E. A. Dalmasso, A. Wirthlin, N. Pfeiffer, SELDI-TOF-MS proteinchip array profiling of tears from patients with dry eye, Investigative ophthalmology & visual science 46 (3) (2005) 863–876. doi:https://doi.org/10.1167/iovs.04-0448.

[71] F.-H. Grus, A. J. Augustin, Analysis of tear protein patterns by a neural network as a diagnostical tool for the detection of dry eyes, ELECTROPHORESIS: An International Journal 20 (4-5) (1999) 875–880. doi:https://doi.org/10.1002/(SICI)1522-2683(19990101)20:4/5<875::AID-ELPS875>3.0.CO;2-V.

[72] F. Grus, A. Augustin, N. Evangelou, K. Toth-Sagi, Analysis of tear-protein patterns as a diagnostic tool for the detection of dry eyes, European Journal of Ophthalmology 8 (2) (1998) 90–97. doi:http://dx.doi.org/10.1177/112067219800800207.

[73] H. R. Choi, J. H. Lee, H. K. Lee, J. S. Song, H. C. Kim, Association between dyslipidemia and dry eye syndrome among the korean middle-aged population, Cornea 39 (2) (2020) 161–167. doi:http://dx.doi.org/10.1097/ICO.0000000000002133.

[74] S. M. Nam, T. A. Peterson, A. J. Butte, K. Y. Seo, H. W. Han, Explanatory model of dry eye disease using health and nutrition examinations: Machine learning and network-based factor analysis from a national survey, JMIR medical informatics 8 (2) (2020) e16153. doi:http://dx.doi.org/10.2196/16153.
URL http://medinform.jmir.org/2020/2/e16153/

[75] M. Kaido, M. Kawashima, N. Yokoi, M. Fukui, Y. Ichihashi, H. Kato, M. Yamatsuji, M. Nishida, K. Fukagawa, S. Kinoshita, K. Tsubota, Advanced dry eye screening for visual display terminal workers using functional visual acuity measurement: the moriguchi study, British Journal of Ophthalmology 99 (11) (2015) 1488–1492. doi:http://dx.doi.org/10.1136/bjophthalmol-2015-306640.

[76] J.-P. C. Gullion, Tear film structure of the contact lens wearer, Ph.D. thesis, City University, London (1990).

[77] F. J. Holly, Physical chemistry of the normal and disordered tear film, Transactions of the ophthalmological societies of the United Kingdom 104 (Pt 4) (1985) 374–380.

[78] W. Rasband, Imagej, http://imagej.nih.gov/ij/.

[79] M. Cruzat, Andrea, M. Qazi, Yureeda, M. Hamrah, Pedram, In vivo confocal microscopy of corneal nerves in health and disease, The ocular surface 15 (1) (2016) 15–47. doi:https://doi.org/10.1016/j.jtos.2016.09.004.

[80] E. Villani, L. Marelli, A. Dellavalle, M. Serafino, P. Nucci, Latest evidences on meibomian gland dysfunction diagnosis and management, The Ocular Surface 18 (4) (2020) 871–892. doi:https://doi.org/10.1016/j.jtos.2020.09.001.
URL https://www.sciencedirect.com/science/article/pii/S1542012420301415

[81] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60 (6) (2017) 84–90. doi:https://doi.org/10.1145/3065386.

[82] D. A. Dartt, Dysfunctional neural regulation of lacrimal gland secretion and its role in the pathogenesis of dry eye syndromes, The ocular surface 2 (2) (2004) 76–91. doi:https://doi.org/10.1016/S1542-0124(12)70146-5.

[83] K. Wan, L. Chen, A. Young, Depression and anxiety in dry eye disease: a systematic review and meta-analysis, Eye 30 (2016) 1558–1567. doi:https://doi.org/10.1038/eye.2016.186.

[84] Y. A. Ambaw, D. P. Timbadia, M. Raida, F. Torta, M. R. Wenk, L. Tong, Profile of tear lipid mediator as a biomarker of inflammation for meibomian gland dysfunction and ocular surface diseases: Standard operating procedures, The Ocular Surface (2020). doi:https://doi.org/10.1016/j.jtos.2020.09.008.
URL https://www.sciencedirect.com/science/article/pii/S1542012420301488

[85] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., Underspecification presents challenges for credibility in modern machine learning, arXiv preprint arXiv:2011.03395 (2020).

[86] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC medicine 17 (1) (2019) 1–9.

[87] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA) - Protein Structure 405 (2) (1975) 442 – 451. doi:10.1016/0005-2795(75)90109-9.

[88] L. I. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1) (1989) 255–268. doi:https://doi-org.ezproxy.uio.no/10.2307/2532051.
URL https://www.jstor.org/stable/2532051

[89] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174. doi:https://doi.org/10.2307/2529310.
URL http://www.jstor.org/stable/2529310

[90] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica 22 (3) (2012) 276–282.

[91] H. Akoglu, User's guide to correlation coefficients, Turkish Journal of Emergency Medicine 18 (3) (2018) 91–93. doi:https://doi.org/10.1016/j.tjem.2018.08.001.
URL https://www.sciencedirect.com/science/article/pii/S2452247318302164

[92] H. Cramér, Mathematical methods of statistics (1945).

[93] D. S. Moore, Introduction to the practice of statistics (2017).

[94] T. Birsan, D. Tiba, One hundred years since the introduction of the set distance by dimitrie pompeiu, in: IFIP Conference on System Modeling and Optimization, Springer, 2005, pp. 35–39. doi:https://doi.org/10.1007/0-387-33006-2_4.

[95] P. Jaccard, The distribution of the flora in the alpine zone, The New phytologist 11 (2) (1912) 37–50. doi:https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[96] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, A. Sethi, A dataset and a technique for generalized nuclear segmentation for computational pathology, IEEE Transactions on Medical Imaging 36 (7) (2017) 1550–1560. doi:10.1109/TMI.2017.2677499.

[97] A. L. B. Benjamin Kompa, Jasper Snoek, Second opinion needed: communicating uncertainty in medical machine learning, npj Digital Medicine 4 (1) (2021) 1–6. doi:http://dx.doi.org/10.1038/s41746-020-00367-3.

A. Supporting information

A.1. Performance scores used

If there are two categories available, the task is referred to as binary classification, while more than two categories is referred to as multi-class. For binary classification, the true outcome belongs to one of two categories, e.g., healthy or ill, often referred to as positive (P) or negative (N). A binary classifier assigns new data instances to these two categories, and the prediction can be either true (T), meaning correct, or false (F), meaning incorrect. The outcome can then belong to one of the four categories true positive (TP), true negative (TN), false positive (FP) and false negative (FN), and sum to the total number of instances in the data set. From these, we can calculate a variety of performance scores, some of which are listed in Section 2.5. We provide mathematical expression for these below. The remaining performance scores encountered in the

reviewed studies are outlined after.

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A.1}$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{A.2}$$

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

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{A.4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A.5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{A.6}$$

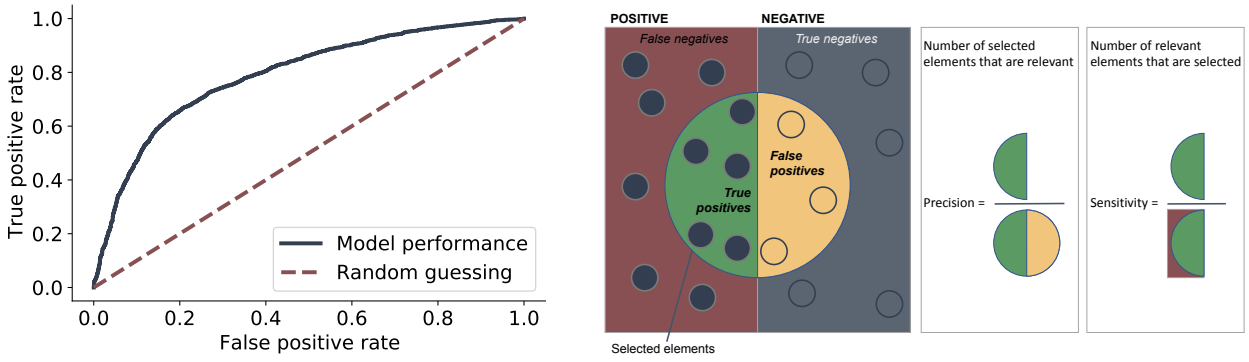
$$\text{F1 score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{A.7}$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{A.8}$$

$$\text{False negative rate} = \frac{\text{FN}}{\text{FN} + \text{TP}}\,. \tag{A.9}$$

Although binary classification tasks involve assigning instances to one of two classes, e.g., 0 and 1, most machine learning classifiers can output the distance of an instance to the decision boundary, i.e., a decimal number in the interval $[0, 1]$. A common interpretation of this number is class probability or classification confidence, meaning that an output close to either number indicates a confident classification, while an output closer to the classification threshold indicates that the classifier is not capable of assigning the instance to a class. The classification threshold is thus the numerical value that separates the two classes, and the confusion matrix entries vary with this threshold. Unless otherwise specified, its value is usually 0.5. Here, we introduce two metrics that can be constructed by varying this threshold from 0 to 1. First, the receiver operating characteristic curve is constructed from the curves of the true and false positive rates obtained by varying the classification threshold. Optimally, the true positive rate is 1 for any threshold, while a classifier which always guesses randomly produces a diagonal line, as shown in Figure A.4a. The AUC value is calculated by summing the area under the receiver operating characteristic curve, and its maximum value is 1.

There is a trade-off between the precision and sensitivity: A high precision minimizes the false positives, which might result in missing positive instances, while a high sensitivity minimizes the false negatives, which can result in a increased number of false alarms. Which one should be prioritised depends on the problem at hand, and a study prioritising or reporting only one of these should argue why. The precision and the sensitivity are visualised in Figure A.4b, which highlights the trade-off between the two. They can be combined into a single number, by plotting them against each other for different classification threshold



(a) (b)

Figure A.4: (a) A receiver operating characteristic curve, and (b) A visual representation of the sensitivity, Equation (A.4), and the precision, Equation (A.5), highlighting the trade-off between the two.

values and calculating the area under the resulting so-called precision-recall curve.

Pearson's correlation coefficient measures the linear correlation between two data sets, and is calculated as

$$\mathrm{r} = \frac{\Sigma(x_i - \overline{x}) \times (y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \times (y_i - \overline{y})^2}}\,, \tag{A.10}$$

where r is the Pearson's correlation coefficient, x_i and y_i are the observed values in each data set and $\overline{x}$ and $\overline{y}$ are the mean values for each data set. The value ranges from -1 to 1, where -1 indicates perfect negative linear correlation and 1 perfect positive linear correlation, while 0 indicates no linear correlation between the data. For binary classification, Pearson's correlation coefficient takes on a simple form, referred to as Matthews correlation coefficient [87]. It measures the correlation between the true and predicted class instances, and ranges from -1 to 1. Here, 0 indicates that the classifier guesses randomly, and 1 and -1 indicate complete agreement and disagreement, respectively, between the model predictions and the true outcome. It can be calculated from the confusion matrix entries as

$$\text{Matthews correlation coefficient} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}\,. \tag{A.11}$$

The *concordance correlation coefficient* measures the agreement between two data sets by measuring the variation around the 45 degrees concordance line through the origin [88]. The value ranges between 1 and -1. When the two data sets share mean and standard deviation, the concordance correlation coefficient equals the Pearsons's correlation coefficient. In all other cases, the concordance correlation coefficient will be lower than the Pearson's correlation coefficient. The value is calculated as

$$\text{Concordance correlation coefficient} = \frac{2 s_{xy}}{{s_x}^2 + {s_y}^2 + (\overline{x} - \overline{y})^2}\,, \tag{A.12}$$

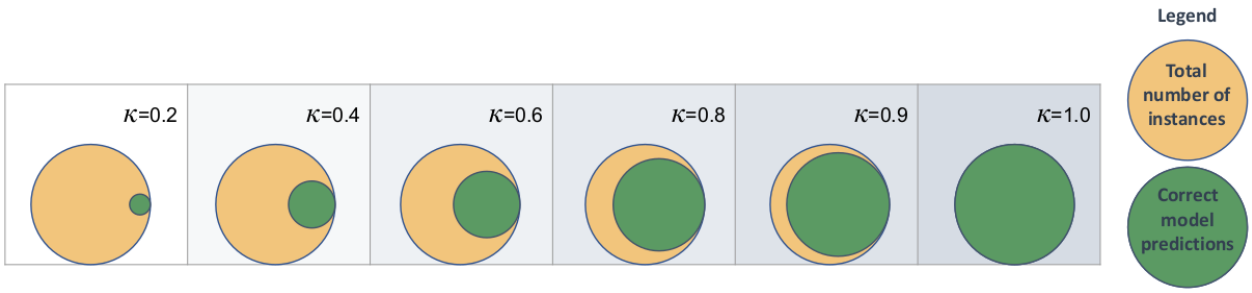


Figure A.5: Kappa values for different degrees of agreement. The illustration is based on [90, Figure 2].

where $\overline{x}$ and $\overline{y}$ are the mean values of the two data sets x and y, ${s_x}^2$ and ${s_y}^2$ are the variances for each data set and ${s_{xy}}^2$ is the covariance between the data sets [88].

Root mean squared error is commonly used for regression problems and represents the difference between the model predictions and the observed values. The value is calculated as

$$\text{Root mean squared error} = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}\,, \tag{A.13}$$

where n is the number of instances in the data set and $\hat{y}_i$ and y_i is the model prediction and observed value for instance i, respectively.

The *Kappa index* measures the agreement between two raters, e.g., the model predictions and labels during classification [89]. It is calculated as

$$\kappa = \frac{p_o - p_e}{1 - p_e}\,, \tag{A.14}$$

where p_o is the observed probability of agreement, which equals the accuracy defined in Equation (A.3), and p_e is the expected probability of agreement due to chance, defined as

$$p_e = \frac{(\text{TP} + \text{FP}) \times (\text{TN} + \text{FN}) \times (\text{FN} + \text{TP}) \times (\text{FP} + \text{TP})}{\text{Total} \times \text{Total}}\,. \tag{A.15}$$

where $Total$ is the total number of instances. The highest possible value is 1, representing perfect agreement, and values above 0.8 are typically regarded as excellent [89]. An illustration of the κ index values for the proportion of correct model predictions is provided in Figure A.5.

Cramér's V measures the association between two categorical variables that belong to more than two categories each. When there are two categories for each variable, Cramér's V equals the φ coefficient [91]. It is calculated via

$$\text{Cramér's V} = \sqrt{\frac{\frac{\chi^2}{n}}{\text{Min(cat1 - 1, cat2 -1)}}}\,, \tag{A.16}$$

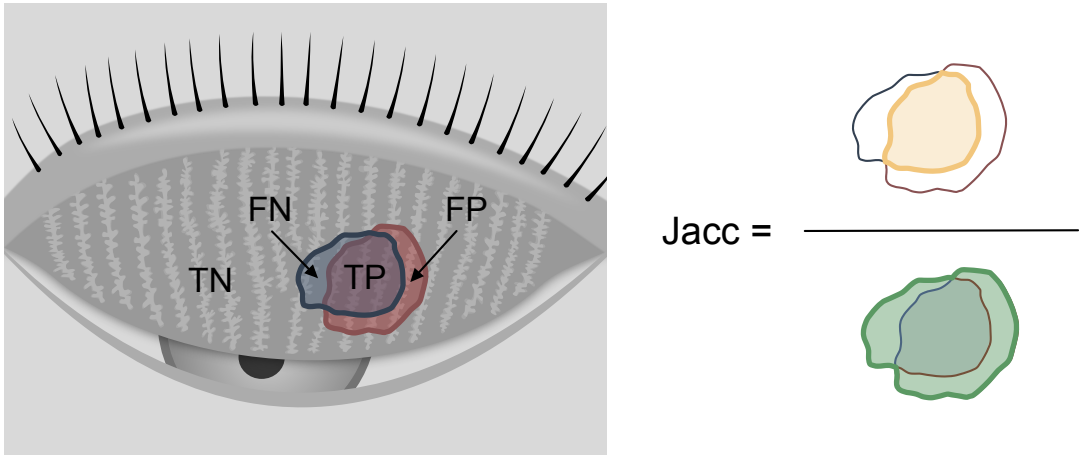


Figure A.6: A visual representation of the Jaccard index. FN = False Negative; TN = True Negative; TP = True Positive; FP = False Positive; Jacc = Jaccard index.

where χ^2 is the usual chi-squared statistic, n is the number of instances, and $cat1$ and $cat2$ are the number of possible categories for each variable. The value ranges from 0 to 1, representing no and perfect correlation between the variables, respectively [92].

In hypothesis testing, the p-value is the probability under a specific model of obtaining test results at least as extreme as those observed, under the assumption that the null hypothesis H_0 is true. H_0 is commonly defined as no difference between two data sets, while the alternate hypothesis H_a states that there is a difference. Consequently, a low p-value indicates that the result is not likely under the null hypothesis, and thus strengthens the belief in H_a [93].

The *Average Pompeiu-Hausdorff distance* reflects the distance between estimated values and true values in a metric space [94]. Lower values imply small differences between the two metric spaces. The Pompieu-Hausdorff distance H between the subsets a and b is calculated via

$$H(a,b) = max(H(a,b), H(b,a))\,. \tag{A.17}$$

The aggregated *Jaccard index* is an extension of the global Jaccard index also used to measure the similarities between two sample sets [95]. A high value indicates small differences between the sample sets. The calculation of the aggregated Jaccard index is described by Kumar et al. [96], and Figure A.6 shows a visualisation. For image segmentation, the *support* for a segmented area can be calculated as the number of pixels in the segmented area divided by the number of background pixels [42].

A.2. Measuring model uncertainty

Uncertainty estimates are useful in order to evaluate how certain a machine learning model is about the predictions. High uncertainty might suggest that a human expert also should have a look at the instance [97]. Among the reviewed studies, some choose not to use the model predictions of DED when the predicted probabilities are too close to 0.5, reflecting that the model is uncertain [75]. Others report the standard deviation of the model performance scores [13, 36, 37, 64, 34, 35, 42, 49, 63]. Some computes the confidence intervals for the model performance scores [32, 39, 74, 63, 63]. A comprehensive discussion about quantifying uncertainty for medical machine learning models can be found in [97].