Deep Learning Based Unsupervised and Semi-supervised Classification for Keratoconus*

1st Nicole Hallett
The Sydney Eye Hospital
The University of Sydney
Sydney, Australia
nhal6992@uni.sydney.edu.au

2nd Kai Yi
School of Mathematics and Statistics
The University of New South Wales
Sydney, Australia
ykyrie822@gmail.com

3rd Josef Dick
School of Mathematics and Statistics
The University of New South Wales
Sydney, Australia
josef.dick@unsw.edu.au

4th Christopher Hodge
The Sydney Eye Hospital
The University of Sydney
Sydney, Australia
christopher.hodge@vei.com.au

5th Gerard Sutton
The Sydney Eye Hospital
The University of Sydney
Sydney, Australia
gerard.sutton@vei.com.au

6th Yu Guang Wang
School of Mathematics and Statistics
The University of New South Wales
Sydney, Australia
yuguang.wang@unsw.edu.au

7th Jingjing You
The Sydney Eye Hospital
The University of Sydney
Sydney, Australia
jing.you@sydney.edu.au

Abstract—The transparent cornea is the window of the eye, facilitating the entry of light rays and controlling focusing the movement of the light within the eye. The cornea is critical, contributing to 75% of the refractive power of the eye. Keratoconus is a progressive and multifactorial corneal degenerative disease affecting 1 in 2000 individuals worldwide. Currently, there is no cure for keratoconus other than corneal transplantation for advanced stage keratoconus or corneal cross-linking, which can only halt KC progression. The ability to accurately identify subtle KC or KC progression is of vital clinical significance. To date, there has been little consensus on a useful model to classify KC patients, which therefore inhibits the ability to predict disease progression accurately.

In this paper, we utilised machine learning to analyse data from 124 KC patients, including topographical and clinical variables. Both supervised multilayer perceptron and unsupervised variational autoencoder models were used to classify KC patients with reference to the existing Amsler-Krumeich (A-K) classification system. Both methods result in high accuracy, with the unsupervised method showing better performance. The result showed that the unsupervised method with a selection of 29 variables could be a powerful tool to provide an automatic classification tool for clinicians. These outcomes provide a platform for additional analysis for the progression and treatment of keratoconus.

Index Terms—Variational Autoencoder, Multilayer Perceptron, Cornea, Keratoconus, Bayesian Neural Networks, Clustering, Deep Learning, Semi-supervised Learning, Dimensionality Reduction, Diagnosis, Amsler-Krumeich Classification

I. INTRODUCTION

Artificial intelligence (AI) is an emerging field in ophthalmic science and medicine. While many researchers have utilised deep and machine learning to monitor and analyse ocular diseases such as diabetic retinopathy [1], macular degeneration [2] and glaucoma [3], to date little AI research has been done within the study of human corneal disease.

Keratoconus (KC) is a bilateral, asymmetric, progressive corneal disease, affecting some 1 in 2,000 patients worldwide (Pearson et al. 2000). It is characterised by central and para-central corneal thinning, leading to induced myopia and irregular astigmatism causing deterioration of the patient's best-corrected visual acuity [4]. Currently, there is little consensus on the etiology of KC however, both genetic and environmental risk factors are considered to play a role [5]. Research suggests a complex matrix of risk factors including gender, age, atopy, sun exposure, geography, allergies, eye rubbing, contact lens wear, dominant sleeping side, body mass index, amongst others. Outside eye rubbing, however, the presence of other factors have provided contradicting findings within the literature [6]–[8].

The main diagnostic tool currently utilised in KC is corneal topography and tomography, which describes the surface curvature of the cornea. Various methods to obtain topographical values exist, including Placido ring and Scheimpflug camera measurement. The methodology impacts the available metrics. Supplementary diagnostic technology may be used including Optical coherence tomography (OCT), which utilises low coherence interferometry to produce a two or three-dimensional image. This may provide additional quantitative and qualitative information on aspects such as corneal thickness and both anterior chamber angle and depth information [9]. Although topography and tomography represent a key diagnostic indication, it is noted that a single diagnostic factor is not sufficiently accurate to confirm an early diagnosis or indeed, disease progression [10].

Corneal transplantation replaces the diseased cornea with donor tissue and may restore corneal regularity and best-corrected visual acuity. However, the surgery and postoperative process is not without significant risks and considered a final treatment option only. The most effective intervention currently available to halt the progression of KC is corneal cross-linking (CXL). Cross-linking is indicated in the presence of disease progression, identified by a combination of increasing corneal curvature, irregularity and refractive changes. The decision to

proceed to CXL represents a clinical challenge as patients will progress at different rates.

Similarly, not all patients will progress to require corneal transplantation for visual rehabilitation. Although routinely valid, surgical complications have been identified and CXL does not appear to work in between 5-20% of procedures [11]. Projecting both the likelihood and rate of progression against the risks of surgical intervention is, therefore, key to optimising the timing of CXL intervention and the potential refractive and visual outcomes. Consequently, an improved model of progression prediction may prove invaluable.

To achieve an accurate model of prediction, it remains essential to classify patients and the stages of the disease process. Currently, no universal classification scheme exists for keratoconus [12]. The purpose of this pilot study, which is the first step in developing an accurate model of disease progression prediction, is to accurately classify KC patients through the use of machine learning.

This research aims to examine machine learning models through the use of patient data to develop a model that can accurately classify different stages of KC, with two key objectives:

- To use supervised and unsupervised models to classify different stages of KC;
- To compare the accuracy of the supervised and unsupervised models in the classification of KC.

The paper is organised as follows. In Section II, we review the related work on Keratoconus diagnosis by machine learning methods. Section III describes a keratoconus patient data set collected in a private ophthalmic clinic (Vision Eye Institute (VEI) Chatswood), which is used in this study. In Section III-B, we develop a variational autoencoder (VAE) with Gaussian mixture classifier. The VAE clusters the corneal data into four A-K classes to distinguish the severity of KC in our cohort. In Section III-C, we develop a multilayer perceptron model to predict the A-K classification label from known labels. In Section IV, we apply the VAE model and MLP model to the clustering and classification tasks for the keratoconus patient data set. The VAE with application to our corneal data set demonstrates excellent performance with clustering accuracy. The MLP has state of the art performance on the real corneal data. In Section V, we outline the results of the study, as well as the next steps for this research.

II. LITERATURE REVIEW

Deep learning has significantly improved object recognition and medical image analysis [13]. In recent years there has been an increase in research with deep learning and ocular disease.

The aim of machine learning in KC research to date has predominantly been to identify early or subclinical forms of the disease. A recent review by Lin et al. [14] details 17 publications concerning the detection of KC, 15 of the 17 reviewed papers utilised topographic maps as the primary input, with 65% of the papers seeking to classify KC patients, and 30% seeking to distinguish KC or sub-KC. While some of the reviewed papers utilised multiple machine learning models, 53%

utilised neural networks, 23% utilised discriminant analysis. Differentiation across accuracy, sensitivity and specificity was apparent in the reviewed papers with ranges reported between 65.2%-100%, 63%-100% and 82%-100% respectively.

Of most relevance to this study, several cohorts utilised the Pentacam corneal topography and tomography unit, thereby allowing objective comparisons. See [15]–[17]. Further, Yousefi et al. [17] used an unsupervised algorithm to cluster KC associated variables with the greatest accuracy effectively.

Kovacs et al. [16] undertook a study with a Pentacam HR Scheimpflug device. This retrospective case-control study utilised a multilayer perceptron neural network to assess the corneal symmetry in 60 eyes from 30 patients who had unilateral keratoconus. This research classified patients based on videokeratography and clinical signs through the framework of KISA, the three components of which are central keratometry (K), the inferior-superior (I-S) value and keratoconus percentage index [12]. The accuracy of machine learning classification in this study was based on the Pentacam progress index (PPI). However, while this functionality is readily available on Pentacam devices, it has not widely been used to classify KC within the existing literature.

Kovacs et al. used classifiers trained on variables through the creation of bilateral data for each parameter. The best neural network architecture was determined from a feedforward network based on highest accuracy, through the training and test sets of 70% and 30% respectively. Results indicated the highest accuracy was to identify subtle corneal changes in unilateral KC patient's control eyes which have both sensitivity and specificity at 90%. However, while the study demonstrated high accuracy between normal, subnormal and KC groups, it did not discriminate between severity levels of KC within the KC group. Further, as the control group were represented by the opposite and previously non-diagnosed eye, the practical benefit of these findings remains unclear. KC is considered bilateral disease albeit often highly asymmetrical in the presentation. Confirmation of a diagnosis of KC in the less affected eye is therefore to be expected.

Hwang et al. [15] utilised multivariate logistic regression analysis and hierarchical algorithm to determine the optimal objective, machine derived variables and combinations, utilising an approach of combining metrics from two devices Pentacam and Spectral Domain OCT. This retrospective case-control study analysed asymmetric clinically normal fellow eyes from 30 KC patients and 60 clinically normal eyes from 60 bilaterally normal control patients. While the authors did not specify training and testing set sizes, in establishing sensitivity, specificity and area under curve (AUC), it demonstrated the ability to reduce the 24 variables to 13, utilising the hierarchical clustering method, highlighting the benefits of utilising multiple devices and inputs. Results were highest for the clustered combination from both the Pentacam and OCT device, with the combined 13 variables as listed in Table I below resulting in 100% sensitivity and 100% specificity distinguishing KC from control corneas.

Yousefi et al. [17] utilised unsupervised machine learning

Pentacam and OCT combined	Resultant 13 variables
	Temporal Outer
	Pachymetry Minimum
	Temporal Inner
	Index Vertical Asymmetry
	Central Thickness
	Epithelium Standard deviation
	Minimum - Medium SD-OCT
	Index Surface Variance
	Inferior Temporal Inner
	Epithelium Minimum-maximum
	Minimum SD-OCT
	Superior Outer
	Superior Inner

to predict KC severity. The study utilised Swept-Source OCT images (CASIA), to analyse 3,000 eyes representing a significant data set. Unlike most studies the Ectasia Severity Index (ESI) was used to determine participant suitability. The ESI, which analyses corneal changes and degeneration is an instrument guided screening index, is not routinely used in clinics minimising the practical impact of the research.

This study utilised Swept-Source OCT images (Casia) of 12,242 eyes from multiple centres in Japan, from which 3,156 eyes with valid ESI were selected. Of interest, the paper only outlines patient selection based on ESI and does not include details on patient history. The algorithm is comprised of 3 key steps, principal component analysis (PCA) to reduce input data from 420 to 8 significant components linearly; manifold learning to reduce non-linearly parameters to eigen-parameters; density-based clustering to identify keratoconic eyes.

This study utilised t-SNE to consider clustering of variables within the analysis, resulting in four different clusters of patients, reflecting the classification of patients. The study achieved a specificity and sensitivity of identifying healthy eyes from KC eyes of 94.1% and 97.7% respectively. The results of this study distinguished KC from normal eyes, rather than establishing a well-defined model of stages within KC.

While emerging, most of the machine learning studies with KC have been undertaken, focusing primarily on the classification of pachymetric images. To date, machine learning has not been utilised to establish an association between the multifactorial variables of KC.

In the present study, Pentacam image data combined with risk factors such as age, gender, eye rubbing and others were included. Both supervised, and unsupervised machine learning was used and compared to a widely used classification system: Amsler-Krumeich (A-K) classification system. The primary goal is to accurately classify KC patients, to develop a more accurate model for early disease progression prediction, rather than identifying or classifying KC from subclinical and normal samples.

III. MATERIALS AND METHODS

A. Patient corneal data set

This is a retrospective single centre study approved by the University of Sydney Human Research Ethics Committee (HREC 2013/1041). Patient data from 124 KC patients was collected from Vision Eye Institute Chatswood between 2014-2017. The medical records of each patient were reviewed and analysed, of which 79% (63.7%) were male. The information in Table II was extracted to be utilised as variables of consideration in both supervised and unsupervised models.

TABLE II
VARIABLES UTILISED IN STUDY

Variable	Source
Gender	Patient questionnaire
Age	Patient questionnaire
Nationality	Patient questionnaire
Diabetes	Medical record
Atopy	Medical record
Allergy	Medical record
Hypertension	Medical record
Other disease presence	Medical record
Length of time since diagnosis	Patient questionnaire
Known eye history	Patient questionnaire
Family history	Patient questionnaire
Eye rubbing	Patient questionnaire
Primary optical aid	Patient questionnaire
Uncorrected distance visual acuity	VEI clinician assessment
Corrected distance visual acuity	VEI clinician assessment
Presence of hydrops	VEI clinician assessment
Corneal scarring	VEI clinician assessment
Vogt's Striae	VEI clinician assessment
Fleischer's ring	VEI clinician assessment
Refractive sphere	Pentacam
Refractive cylinder	Pentacam
Refractive axis	Pentacam
Flat keratometry	Pentacam
Steep keratometry	Pentacam
Thinnest pachymetry	Pentacam
Location X axis	Pentacam
Location y axis	Pentacam
Central pachymetry	Pentacam
Amsler-Krumeich (AK) classification	VEI clinician

All patients were classified, through clinician experience, based on the Amsler-Krumeich (A-K) classification, which is comprised of mean-K readings on the anterior curvature sagittal map, thickness at the thinnest location and the refractive error of the patient, and biomicroscopy. Table III [18] shows how A-K classification groups patients.

B. Clustering for corneal data by variational autoencoder

Variational autoencoder (VAE) [19] is a Bayesian deep neural network, which consists of an encoder and a decoder and a latent variable layer. The encoder and decoder which are deep neural networks are used to extract features from the input data and to generate the same type of output data from the latent features respectively. The encoder is a deep net built to learn features of the input data which are then passed to the latent variable layer. In the decoder, the latent features are used to generate the output data with the same format of the original

Grad	Characteristic
1	Eccentric steeping Myopia and astigmatism < 5.00D Mean central K readings < 48.00D
2	Myopia and astigmatism 5.00-8.00D Mean central K readings < 53.00D Absence of scarring Minimum corneal thickness $> 400\mu m$
3	Myopia and astigmatism 8.00-10.00D Mean central K readings > 53.00D Absence of scarring Minimum corneal thickness 300-$400\mu m$
4	Refraction not measurable Mean central K readings > 55.00D Central corneal scarring Minimum corneal thickness 200μm

data. Between the encoder and decoder, the latent variable layer uses Gaussian random sampling to generate latent features. The clustering is then obtained by features from the latent variable layer between the encoder and decoder.

In this work, we take the encoder and decoder as multilayer perceptron (MLP) models, which is a fully connected deep neural network. The input data was trained by the VAE of the above network architecture (see Figure 1). The encoder uses a deep net to compress each high-dimensional input sample into a two-dimensional real vector. By this, the encoder extracts the features of the input and is then clustered by a Gaussian mixture model [20] to a given number of classes.
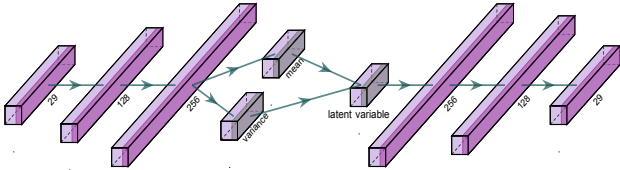

Fig. 1. VAE architecture: encoder, decoder and latent variable layer; the encoder and decoder are MLP with 2 hidden layers.

Figure 1 shows the VAE model we use in the experiments. The encoder and decoder are MLP with 2-hidden layers. The network architecture of encoder and decoder are 29-128-256-2 and 2-256-128-29, where the 29 is the number of the variables in the corneal data, 256 and 128 are the numbers of hidden neurons of the deep nets. These are hyper-parameters which have been tuned to optimize the performance of the deep network. When the VAE model has been trained, a Gaussian mixture model is applied to cluster the compressed 2D vectors.

The VAE is interpreted as a Bayesian inference model. The N input data of encoder $X = \{x^{(i)}\}_{i=1}^{N}$ and N outputs of decoder $Y = \{y^{(i)}\}_{i=1}^{N}$ and latent feature z are regarded as random variables which are in the probability spaces $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$. The encoder is characterised by the conditional probability

density (CPD) $q_\phi(z|x)$ satisfying

$$p(z) = \int_{\mathcal{X}} q_\phi(z|x) p_X(x)\mathrm{d}x,$$

where p_X is the (unknown) probability density function of the input data. The ϕ signifies the parameters of the encoder neural network.

The decoder is depicted by the CPD $p_\theta(y|z)$, which links the latent variable z by

$$p_\theta(y) = \int_{\mathcal{Z}} p_\theta(y|z) p(z)\mathrm{d}z \quad \forall y \in Y,$$

where p is the probability density function of the latent variable and θ signifies the parameters of the decoder neural network.

The marginal log-likelihood of output data set Y is given by $\log p_\theta(Y) := \sum_{i=1}^{N} \log p_\theta(y^{(i)})$. For sample $i = 1, \ldots, N$, the log-likelihood $\log p_\theta(y^{(i)})$ has a variational lower bound [21]

$$\log p_\theta\left(y^{(i)}\right) \geq \mathbf{E}_z\left[\log p_\theta\left(y^{(i)}|z\right)\right]$$
$$- D_{\mathrm{KL}}\left(q_\phi\left(\cdot|x^{(i)}\right) \| p(\cdot)\right),$$

where the D_{KL} term represents the Kullback-Leibler divergence [22] of the posterior distribution of the output of the encoder and the prior distribution of the latent variable, and the $-\mathbf{E}_z$ term is the reconstruction error of the decoder output. From this, we can define the loss function of the VAE as

$$\mathrm{LOSS} = \sum_{i=1}^{N} \Bigg\{ D_{\mathrm{KL}}\left(q_\phi(\cdot|x^{(i)}) \| p_\theta(\cdot)\right)$$
$$- \mathbf{E}_z\left[\log p_\theta(y^{(i)}|z)\right] \Bigg\}.$$

In training the neural network, the VAE uses back-propagation to minimize the loss function. The $q_\phi(\cdot|x)$ and $p_\theta(\cdot)$ are set as normally distributed density functions. The mean and variance of $q_\phi(\cdot|x)$ are parameters by a deep neural network, and the $p_\theta(\cdot)$ is the standard normal distribution density by which the latent variable z is sampled.

C. Deep neural networks for classification

We use the multi-layer perceptron (MLP) model for corneal data classification. The MLP model is a fully connected deep neural network with multiple layers for a semi-supervised learning task, see [23], [24].

An MLP of depth K is a multivariate function in the form of

$$x_i^{(j+1)} = g_{ij}(x^{(j)} \cdot w_{ij} + b_{ij}),$$

for $j = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, N_j$, where the $x^{(j)} = (x_1^{(j)}, \ldots, x_{N_j}^{(j)})$ is the input of jth layer with N_j neurons, and g_{ij} is the activation function on $\mathbb{R}$ which connects the input and the ith neuron in the jth layer, and w_{ij} and b_{ij} are the trainable weights and biases. For classification task, the last MLP is trained by back-propagation using a stochastic gradient descent optimisation strategy.
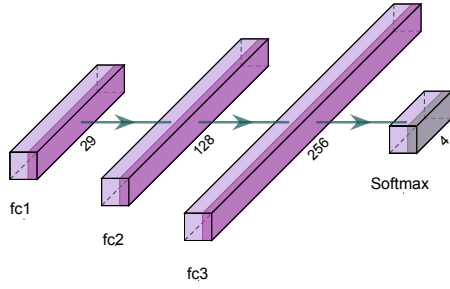
Fig. 2. Network architecture of MLP with three hidden layers. It has 128, 256 and 4 neurons in the second, third and fourth hidden layers.

Figure 2 shows the MLP model we use for corneal data with three hidden layers. The network architecture is 29-128-256-4, with 29 the dimension of the input data (i.e. the number of the variables of corneal data) and $128, 256$ and 4 the numbers of neurons in the hidden layers in turn.

IV. EXPERIMENTAL RESULTS FOR CORNEAL DATA

A. Results and clinical interpretation of VAE clustering

As mentioned, we use VAE to reduce the dimension of input data from the 29 initial variables to 2. We then utilise Gaussian mixture modelling for clustering to classify the compressed sample data into one of the 4 A-K classes. Figure 3 shows the plot of latent 2D feature vectors of the corneal input data. The class label is determined by the AK classification label of the ground truth. We can observe that there is already an embryonic form of clustering, but the classification is not clear as different clusters have many overlaps.

We thus send the latent 2D feature vectors to the Gaussian mixture model, which then clusters the samples into 4 classes. As depicted in Figure 4, the classes of 1,2,3 and 4 relate to A-K classifications of 1-4 as outlined in the introduction. Patients in classes 1 and 4 are independent, as the groups have different mean and variance in R^2. The results also indicate distinct features between classes 1 and 3. Our results infer that classes 1 and 2 are the most similar, as evidenced with overlapping on the cluster plot, which is brought about by miss-clustering of the data for patients classified within groups 1 and 2. The accuracy of the VAE in Figure 4 compared with the ground truth is as high as 80.3%. In 20 repetitions of the test, comparing these results to the ground truth of A-K classification through clinical diagnosis represents a significant outcome with accuracy at 76.9% with Std Dev. of 3%, and the highest accuracy level of 82.4%. The Std Dev. 3%, which is small, reflects the uncertainty in the sampling of the latent variable and Gaussian mixture clustering of the VAE.

Figure 5 shows the ROC curves and AUC for the VAE classifier. The ROC curves for all classes are close to the upper left corner, and the AUC for classes 1,2,3 and 4 achieves high values at 0.91, 0.87, 0.79 and 0.99. It illustrates that the VAE classifier has excellent performance for the corneal data clustering. Meanwhile, the VAE has helped to classify patients
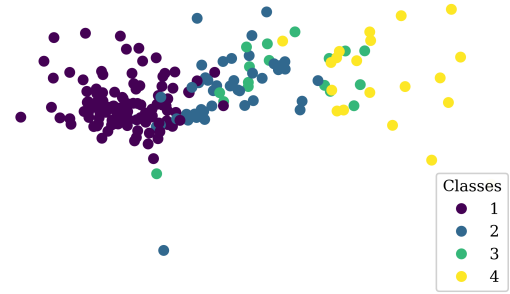

Fig. 3. The corneal data is compressed into two dimensional vector by VAE. They are the input of the Gaussian mixture model for clustering. The 2D vectors are the extracted features of the input data which can then be used for clustering. Here, the class label for each sample is from the ground truth.
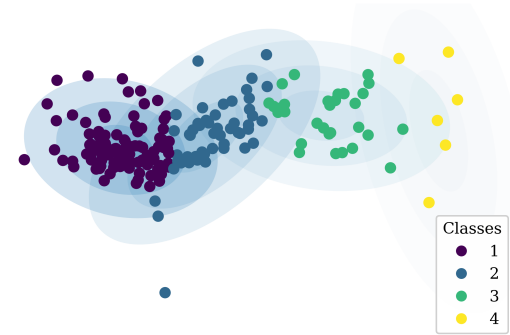

Fig. 4. Corneal data is clustered into 4 classes by Gaussian mixture model using the latent 2D vectors from the VAE model. The clustering accuracy compared to the ground truth is 80.3%. The shallow blue ellipses show the confidence region of the centre of each cluster. The difference between class 1 and the classes 3 and 4 are apparent; the 1st and the 2nd classes are close to each other.

into 4 classes that correspond to the A-K classification. As this is an unsupervised model, it will facilitate clinicians being able to accurately group patients with an A-K classification based on the input 29 variables of real measurement.

B. Results and clinical interpretation of MLP classifier

The MLP we utilise is a three-layer fully connected neural network which learns the diagnosis of A-K classification from the training data. The trained MLP model can be used to find the A-K classification label of the test data set. Of totally 237 samples, we use 64% and 16% in training and validation and 20% in the test. We ran the experiments for 100 repetitions to reduce the randomness due to random shuffle of the data set and model training. We use 100 epochs for each training. Our MLP model demonstrated a mean validation accuracy of 86.37% with epoch and variance of validation of 79-94% across the analysis. The validation loss is decaying fast with a small variation. The mean test accuracy is 75.85% with Std Dev 5.88%. Figure 8 shows the ROC curves for the MLP model
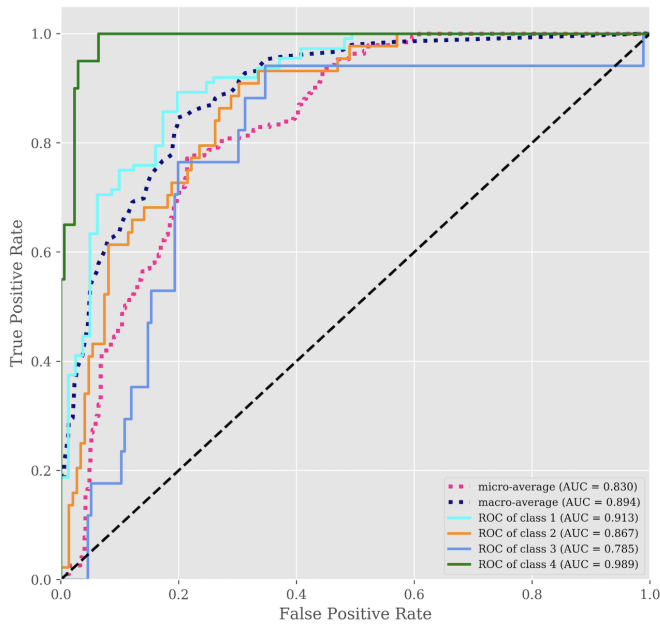
Fig. 5. ROC curves for VAE classifier. The VAE classifier shows excellent performance on each of four classes, with AUC between 0.79 to 0.99.
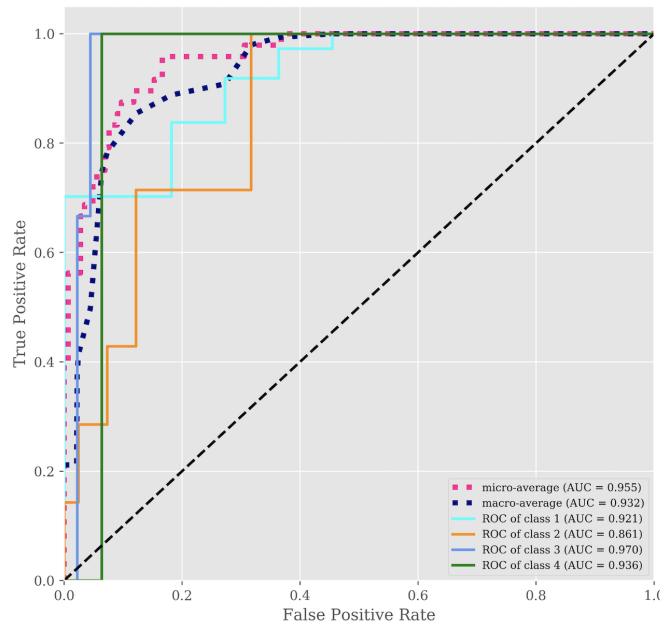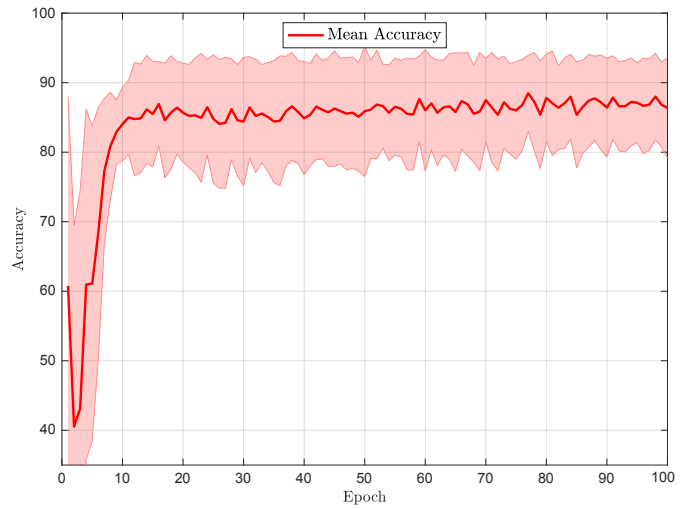

Fig. 7. Validation accuracy of MLP for corneal data; the red curve shows the mean validation accuracy with epoch up to 100 in training; the shallow red region shows the variance of validation accuracy with epoch.
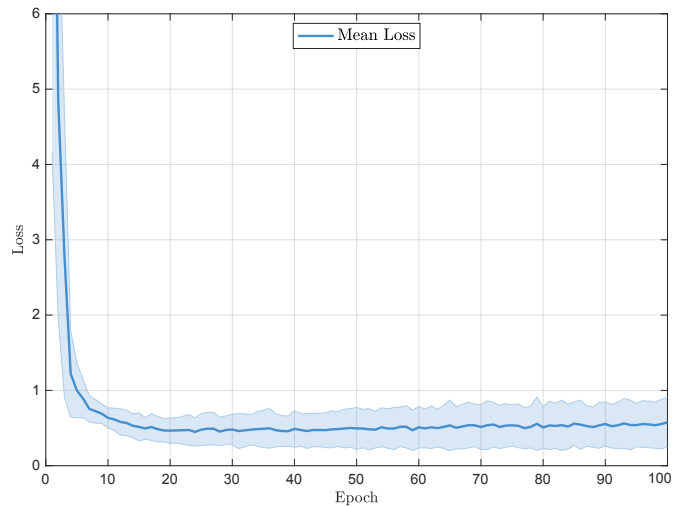

Fig. 8. Validation loss of MLP for corneal data; the blue curve shows the mean validation loss with epoch up to 100 in training; the shallow blue region shows the variance of validation loss with epoch.

on each of four classes on corneal data with AUC values 0.92, 0.86, 0.97 and 0.94. The micro-average and macro-average AUC values (which reflect the general performance of the MLP classifier) are high at 0.96 and 0.93. The ROC curves are all close to the upper left corner, which indicates the excellent performance of the MLP classifier. We can thus infer that the MLP model demonstrates promising potential as an accurate artificial intelligence diagnosis for corneal diseases.

V. CONCLUSION AND DISCUSSION

We utilised an MLP model to learn the corneal disease diagnosis from labelled training data, and then developed a Bayesian neural network (Variational Autoencoder plus a Gaussian mixture model) to determine the degree of corneal disease from unlabelled corneal data. Both methods achieve a state of the art performance on real corneal data.

Fig. 6. ROC curves for MLP classifier. The MLP classifier shows excellent performance for each of four classes, with AUC between 0.86 and 0.97.

In this study, the unsupervised VAE model resulted in higher accuracy than the MLP model. Training against an existing classification system, that is the A-K classification system in this study, is not needed with the unsupervised model. There are already multiple KC classification systems [16], [17] utilised in research and clinical settings, resulting in inconsistency through research output and clinical interpretation. This VAE model with the 29-dimensional input could thus represent a potential independent and standardised classification system for KC.

As VAE results (see Figure 4) showed, from left to right, 4 classes distinguished the patients from early to late stages, which fits well with the progression pattern of the disease. The cluster centre between classes 1 and 2 was closer to each other, suggesting less differentiation between features and variables of these groups. It is consistent with the clinical observation that early stages of KC are hard to distinguish. Furthermore, this clustering map seems to not only classify patients similarly to the A-K system, but it also visualises how much closer the features of patients (represented by the dots) are towards the next stage. Identifying where a patient may sit within their classification, that is at the beginning, middle or upper extremity may have significant additional impact contributing to the analysis of progression.

Accordingly, the next step in our research is to expand the model to include control samples inclusive of longitudinal outcomes. Unlike research to date, a vital goal of this study has been not only to identify whether a patient has KC or not, but rather to be able to accurately group patients within the framework of the existing A-K KC classification stages.

Also of significance, the difference between this study and those reviewed is that we looked at both eyes of KC patients, as compared to the studies [16] and [15] which sought to identify sub-clinical KC in the fellow eyes of unilateral KC patients.

This study is to be able to accurately group patients within the framework of the existing A-K KC classification stages. We focus on clinician involvement in determining the variables considered and classification, which is an essential difference between our study and those of [15]–[17], which use built-in automated algorithms.

While this study did not achieve the same sensitivity and specificity as the unsupervised model in [17], it is seeking to distinguish different stages within KC, whereas the high accuracy generated in [17] was to differentiate between KC and control. Besides, our study represents a small data set compared to [17].

This study was able to achieve mean accuracy levels of 76% and 80% for supervised and unsupervised models respectively, and we expect that the inclusion of control samples alongside the existing clinical data may lead to additional improvement in these outcomes. It is important to continue training and testing the models to develop an approach which assists clinicians better manage KC and predict disease progression.

REFERENCES

[1] R. C. Date, S. J. Jesudasen, C. Y. Weng *et al.*, "Applications of deep learning and artificial intelligence in retina," *International Ophthalmology Clinics*, vol. 59, no. 1, pp. 39–57, 2019.

[2] N. Motozawa, G. An, S. Takagi, S. Kitahata, M. Mandai, Y. Hirami, H. Yokota, M. Akiba, A. Tsujikawa, M. Takahashi *et al.*, "Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes," *Ophthalmology and Therapy*, pp. 1–13, 2019.

[3] J. M. Ahn, S. Kim, K.-S. Ahn, S.-H. Cho, K. B. Lee, and U. S. Kim, "A deep learning model for the detection of both advanced and early glaucoma using fundus photography," *PLOS One*, vol. 13, no. 11, p. e0207982, 2018.

[4] H. Serdarogullari, M. Tetikoglu, H. Karahan, F. Altin, and M. Elcioglu, "Prevalence of keratoconus and subclinical keratoconus in subjects with astigmatism using Pentacam derived parameters," *Journal of Ophthalmic & Vision Research*, vol. 8, no. 3, p. 213, 2013.

[5] V. M. Tur, C. MacGregor, R. Jayaswal, D. O'Brart, and N. Maycock, "A review of keratoconus: diagnosis, pathophysiology, and genetics," *Survey of Ophthalmology*, vol. 62, no. 6, pp. 770–783, 2017.

[6] A. Davidson, S. Hayes, A. Hardcastle, and S. Tuft, "The pathogenesis of keratoconus," *Eye*, vol. 28, no. 2, p. 189, 2014.

[7] C. W. McMonnies, "Inflammation and keratoconus," *Optometry and Vision Science*, vol. 92, no. 2, pp. e35–e41, 2015.

[8] A. Gordon-Shaag, M. Millodot, I. Kaiserman, T. Sela, G. Barnett Itzhaki, Y. Zerbib, E. Matityahu, S. Shkedi, S. Miroshnichenko, and E. Shneor, "Risk factors for keratoconus in Israel: a case–control study," *Ophthalmic and Physiological Optics*, vol. 35, no. 6, pp. 673–681, 2015.

[9] H. Li, V. Jhanji, S. Dorairaj, A. Liu, D. S. Lam, and C. K. Leung, "Anterior segment optical coherence tomography and its clinical applications in glaucoma," *Journal of Current Glaucoma Practice*, vol. 6, no. 2, p. 68, 2012.

[10] A. Martínez-Abad and D. P. Piñero, "New perspectives on the detection and progression of keratoconus," *Journal of Cataract & Refractive Surgery*, vol. 43, no. 9, pp. 1213–1227, 2017.

[11] M. D. Ozer, M. Batur, S. Mesen, S. Tekin, and E. Seven, "Long-term results of accelerated corneal cross-linking in adolescent patients with keratoconus," *Cornea*, vol. 38, no. 8, pp. 992–997, 2019.

[12] X. Li, H. Yang, and Y. S. Rabinowitz, "Keratoconus: classification scheme based on videokeratography and clinical signs," *Journal of Cataract & Refractive Surgery*, vol. 35, no. 9, pp. 1597–1603, 2009.

[13] J. L. Alio *et al.*, "Keratoconus," *Recent Advances in Diagnosis and Treatment*, 2017.

[14] S. R. Lin, J. G. Ladas, G. G. Bahadur, S. Al-Hashimi, and R. Pineda, "A review of machine learning techniques for keratoconus detection and refractive surgery screening," in *Seminars in Ophthalmology*. Taylor & Francis, 2019, pp. 1–9.

[15] E. S. Hwang, C. E. Perez-Straziota, S. W. Kim, M. R. Santhiago, and J. B. Randleman, "Distinguishing highly asymmetric keratoconus eyes using combined Scheimpflug and spectral-domain OCT analysis," *Ophthalmology*, vol. 125, no. 12, pp. 1862–1871, 2018.

[16] I. Kovács, K. Miháltz, K. Kránitz, É. Juhász, Á. Takács, L. Dienes, R. Gergely, and Z. Z. Nagy, "Accuracy of machine learning classifiers using bilateral data from a Scheimpflug camera for identifying eyes with preclinical signs of keratoconus," *Journal of Cataract & Refractive Surgery*, vol. 42, no. 2, pp. 275–283, 2016.

[17] S. Yousefi, E. Yousefi, H. Takahashi, T. Hayashi, H. Tampo, S. Inoda, Y. Arai, and P. Asbell, "Keratoconus severity identification using unsupervised machine learning," *PLOS One*, vol. 13, no. 11, p. e0205998, 2018.

[18] K. Kamiya, R. Ishii, K. Shimizu, and A. Igarashi, "Evaluation of corneal elevation, pachymetry and keratometry in keratoconic eyes with respect to the stage of Amsler-Krumeich classification," *British Journal of Ophthalmology*, vol. 98, no. 4, pp. 459–463, 2014.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *ICLR*, 2014.

[20] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with Gaussian mixture variational autoencoders," 2016.

[21] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[22] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.