**End-to-End Deep Learning Vector**

**Autoregressive Prognostic Models to Predict**

**Disease Progression with Uneven Time Intervals**

No Author Given

No Institute Given

**Abstract.** We propose an end-to-end deep learning method combining implicit feature extraction and an autoregressive model to predict the future course of a disease or condition. By merging the feature extraction and autoregression into one deep learning model, we can simultaneously train both models together. Our novel approach begins by fine-tuning a pretrained convolutional neural network to extract features from previously obtained images of patients. A trainable autoregression mechanism then predicts the features of the future image and a fully connected layer gives a prognosis based on the predicted features. We utilize a novel time interval scaling, allowing the model to account for uneven time intervals and allowing us to choose the final time point that we wish to predict. Experiments on the Age-Related Eye Disease Study give a testing area under the receiver operating characteristic curve, sensitivity, and specificity of 0.966 (95% CI: 0.947, 0.984), 0.878 (95% CI: 0.810, 0.945), and

0.930 (95% CI: 0.914, 0.947), respectively.

**Keywords:** Prognosis · Autoregressive · Deep Learning

# Introduction

Prognostic models have been developed in various fields to predict the future outcome of disease [29]. These models are a vital component of personalized medicine, allowing clinicians to prepare treatments and to allocate resources accordingly. Most developed prognostic models use traditional statistics, which often rely on features or variables being known and obtainable; this is not always the case in imaging data. More recently, deep learning has allowed automated feature extraction, avoiding the necessity to extract those features ourselves [3,27]. Models that utilize deep learning in prognostic models based on images fall into two main categories; explicit feature extraction and implicit feature extraction. The explicit feature extraction methods utilize deep learning to extract known features automatically, saving considerable time. The extracted features can then be used in a traditional statistical model such as logistic regression or linear mixed-effects model in the case of longitudinal images. However, training automatic feature extraction algorithms require expertly annotated images and assume that we know which features to extract. Previously, these explicit feature extraction and classification algorithms have been trained in two steps, meaning that the classification error cannot be backpropagated to the feature extraction. Implicit feature extraction methods feed a single image, usually the last available image, into a convolutional neural network (CNN) and return a probability of progression. These methods allow the network to extract features implicitly. Previously, this approach has not considered multiple longitudinal images and possibly neglects important temporal patterns.

Here, we propose a method combining both implicit feature extraction and an autoregressive model into a single deep learning framework, allowing both parts to be trained simultaneously. In traditional statistics, the time series parameters must be estimated using a method such as ordinary least squares; however, in our method, we use deep learning to learn the parameters. A window function is employed to allow for more uneven time intervals. The proposed method requires no prior annotations and can be used on multiple longitudinal images.

# Related work

Currently, there exist two main approaches to developing deep learning prognostic models using images. Explicit feature extraction combined with a statistical model and implicit feature extraction in an end-to-end model.

## Autoregression

It is essential to highlight the difference in our work compared to previous deep learning models using autoregression. Autoregression has previously been used in deep learning in a different form for density estimation [32] or as a skip mechanism between hidden layers in generative autoencoders [15]. These models are autoregressive in the sense that each output from the previous layer is used as additional inputs in the next layer. In our model, autoregression acts as a layer on its own at the end of the network to predict the future feature vector in a sequence of images.

## Explicit feature extraction

One approach to developing a prognostic model is first to extract the features, either manually or with an automatic algorithm, and then to use these features as variables in a traditional statistical model.

The field of ophthalmology, in particular, has several models taking this approach, using Optical Coherence Tomography (OCT) [18]. OCT is an imaging technique similar to ultrasound, using light instead of sound. Due to the easy identification of individual retinal layers in OCT, retinal layers can be accurately segmented and features, such as volumes, calculated. One of the first such models proposed by de Sisternes et al. [27] used a deep learning algorithm to extract area, volume, height, and reflectivity of drusen from OCT images before a statistical model predicts the future onset of age-related macular degeneration (AMD) based on those features. Banerjee et al. [6] followed the same feature extraction approach, replacing the statistical model with a Recurrent Neural Network (RNN); they found that this method provided improved performance over a random forest classifier.

There are also several prognostic models developed with this approach in neurology and cardiology. Hilario et al. [17] developed a Cox proportional hazards model [11] to predict the progression of diffuse gliomas based on features automatically extracted by commercial software. Arenja et al. [4] also used commercial software to extract features to assess the prognostic value of magnetic resonance imaging in predicting the outcome of patients with heart disease.

While these models often achieve high predictive performance, they all rely on a feature extraction algorithm. These algorithms require a large amount of training data, labeled by experienced experts, which is expensive to produce and not always possible if features are unknown or difficult to quantify. By using two separate models for feature extraction and prediction, the feature extraction network is disconnected from the final classification, meaning the error cannot be backpropagated to improve feature extraction.

## Implicit feature extraction

An alternative to the feature extraction and statistical model method is to use an end-to-end deep learning model. These models combine implicit feature extraction, with a convolutional neural network and classification, with a fully connected layer, into a single model, meaning we no longer need to label the features during training. Usually, the last available image is used as input into these models to predict the future outcome.

As with the previous approach, there are many developed models using this approach in the field of ophthalmology. Yan et al. [36] used InceptionV3 [30], pretrained on Imagenet [26], and combined the feature vector with genotype and phenotype data to be classified with a fully connected layer. Grassman et al. [14] combined six deep learning networks in a random forest ensemble method; they achieved improved results over any of the networks alone. This model classified patients into 13 separate groups, nine risk of advanced AMD groups, three advanced AMD groups, and one ungradable group. Arcadu et al. [3] utilized multiple color fundus fields to predict diabetic retinopathy progression. A separate Inception V3 network [30] was fit to each field of the retina, and a prediction of progression was produced. The final overall prediction was made using a random forest classifier. Many of these fields are not routinely photographed, making this suitable for some applications only.

This approach has also been applied to neurology. Choi et al. [9] used a 3D CNN to predict conversion to Alzheimer’s disease from MRI scans. The CNN method provided increased performance over a voxel-wise feature extraction and support vector machine method, suggesting this approach is preferable to using feature extraction.

More recently, methods able to deal with longitudinal data have been proposed [7]; these enable the network to better model the progression of the disease.

Here, we propose a longitudinal method using vector autoregression (VAR) to predict what the future features may be.

# Method

The proposed model is unique in utilizing a time series model to predict what the future extracted features may be using the previous features; the problem then becomes a simple classification. The features used are implicit and extracted using a CNN.

Given previous images *I*1*,I*2*,...,In* at time points *t*1*,t*2*,...,tn*, we aim to predict the diagnosis *yn*+1 of some future image *In* at time *tn*+1. We also aim to account for uneven time intervals, where *ti*+1-*ti* = *tj*+1 −*tj* does not necessarily hold.

## Overall framework

While previous models aim to predict the future outcome, our model instead aims to predict the future features, which are then classified as progressing or non-progressing. We first fine-tune a pretrained CNN on each image at each time point; this results in a feature vector at each time point. The feature vectors are then multiplied by an interval scaling, which weights images closer to the outcome time as being more informative. We predict the future feature vector using a vector autoregressive model with trainable parameters. Finally, the predicted feature vector is classified with a fully connected layer. Each of these stages is trained in a single deep learning model, such that the classification error is backpropagated to improve the implicit feature extraction. This framework is displayed in Figure 1.

## CNN

We begin reducing each image at each time point image to a single vector representation. For this, we utilize a convolutional neural network (CNN). In our work, we chose InceptionV3 [30], pretrained on Imagenet [26] followed by global average pooling, as this network has been used by previous single time point methods and obtained excellent results [3,5,36]; however, our method may use any CNN of choice. InceptionV3 improves upon previous inception networks with factorized convolutions and heavy regularization to produce an accurate yet computationally efficient network. Previous experiments show excellent classification performance with a low receptive field resolution, suggesting the network is suitable for images with small objects such as those encountered in medical imaging.

The CNN will reduce each image to a feature vector *xi* of length F; these vectors can be treated as extracted features that the algorithm finds useful in making the final classification.



Early/Intermediate

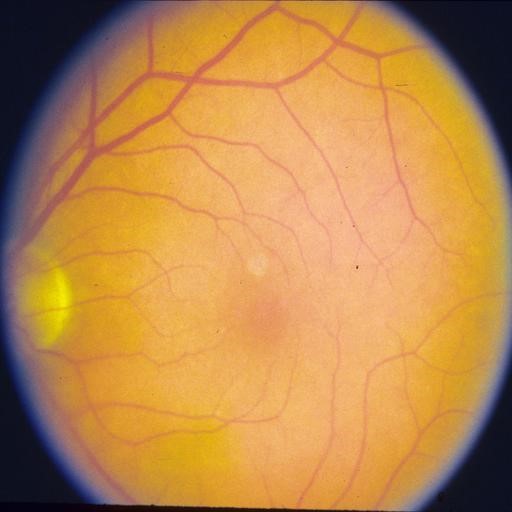
Early/Intermediate



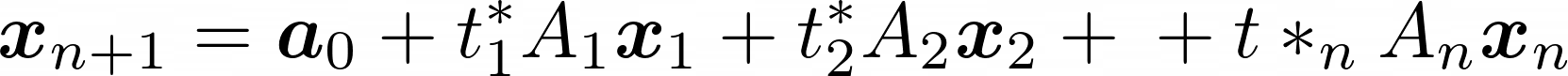
CNN

CNN

…



CNN



Early/Intermediate

Feature Vector

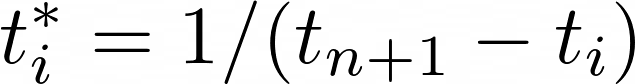
Feature Vector

Feature Vector

Probability

Dense

**Autoregressive model**



**Fig.1.** Framework of the proposed method. The method takes images from multiple time points and extracts a feature vector for each image using a CNN. A future feature vector is then predicted using a Vector Autoregressive (VAR) model. An interval scaling is applied to account for uneven time points. Finally, a dense fully connected, with sigmoid activation, returns the predicted probability of progression. In this example, images of Age-Related Macular Degeneration are shown; however, any disease and imaging modality may be used.

## Time series

To predict the future feature vector *xn*+1 at time *tn*+1, we use time series analysis. Time series theory assumes that each feature is influenced by some random mechanism and is correlated with features at other time points. With these assumptions, we can use the previous features to predict the features of some future image. Here, we outline the time series method utilized in this paper, autoregression (AR), as well as the multivariate generalization, vector autoregression (VAR). These finite parameter models are derived as special cases of the general linear model [23].

**Autoregressive** models perform regression on previous values to predict the next value in the sequence. Autoregression is often used in finance and meteorology to predict stock prices and daily temperatures. The pth order autoregressive model denoted AR(p), is given by:

*xn*+1 = *an*+1 + *anxn* + ··· + *an*−*pxn*−*p* + *et,*

where *xn* is the value we wish to predict using the previous values {*xi*}, {*ai*} are parameters to be determined, and *et* is some stationary random noise.

**VAR** models extend the autoregressive model to the multivariate case by replacing the time series, {*xi*}, with vectors {*xi*} and the parameters, {*ai*}, with matrices {*Ai*}. VAR allows us to consider the relationships between features.

Traditionally the parameters {*Ai*} would be estimated using a method such as ordinary least squares; however, in our method, we create a trainable tensor and use deep learning to learn the parameters instead. Each feature vector, *xi* of length F, obtained from the previous stage, is multiplied by a trainable tensor, *Ai*, of shape *F* × *F*. Finally, the new feature vectors are summed along with an intercept vector *a*0.

## Interval scaling

Often patients may miss appointments or may have their screening intervals relaxed; this can create uneven intervals between visits. These uneven time intervals are accounted for by further multiplying the feature vector by a scalar equal to:

*.*

This also allows us to choose at which time point we wish to predict by changing *tn*+1.

Therefore the final autoregressive model, combined with the interval scaling, to predict *xn* is given by:

*x**.*

## Classification layer

The predicted feature vector *xn*+1 is then classified using the standard dense, fully-connected layer within deep learning. For binary prognosis problems, such as progression/non-progression, the sigmoid activation function may be used, whereas, in multiclass prognosis, the softmax activation is used. When the sigmoid activation is used, the classification layer becomes equivalent to logistic regression.

# Experiments

We demonstrate the proposed method on the Age-Related Eye Disease Study (AREDS) dataset, the most extensive study in Age-Related Macular Degeneration (AMD) [2].

As our proposed model is comprised of two novel concepts, we will consider each of them individually and then combined, to assess the contribution of each. The first model uses a single time point to predict the outcome at the next time point and does not take into account the time between the observation and the prediction; this is similar to previously used methods such as [3,5,36]; however, we do not include additional fields, stereo images, or genetic data, as this information is not routinely available. The second model uses a single time point with the proposed interval scaling method to account for the uneven time intervals. The third model uses vector autoregression (VAR) with multiple time points but does not use the interval scaling. Finally, the proposed model uses both VAR and interval scaling.

## Dataset

AMD is a degenerative eye condition and a leading cause of vision loss worldwide, with a predicted prevalence of 288 million by 2040 [35]. A meta-analysis of 20072009 data estimated that there are 71,000 new cases of late AMD per year in the UK [22]. The current best treatment for AMD is a repeated course of intravitreal antibody injections. The success of treatment depends significantly on early identification and intervention [35].

AMD can be divided into two primary grades, early/intermediate and advanced [2]. Early/intermediate AMD is characterized by small- to medium-sized drusen; these are lipid deposits that form under the retina. Advanced AMD can be further divided into “dry” geographic atrophy (GA) or “wet” neovascular AMD (nAMD); it is possible to have both GA and nAMD.

Images are taken from AREDS, a longitudinal prospective study into AMD. Patients with early/intermediate AMD, between the ages of 55 and 80, were recruited and followed up to 12 years from baseline. Images were taken at time points with varying intervals. From the AREDS dataset, we extracted 4903 eyes, which had three visits displaying early/intermediate AMD and a fourth visit with either early/intermediate or advanced AMD; from these images, we aimed to predict the diagnosis at visit four using the first three images. We split the images into training (60%), validation (20%), and testing (20%) subsets; this is shown in Table 1. Example patient profiles are shown in Figure 2.

## Computation

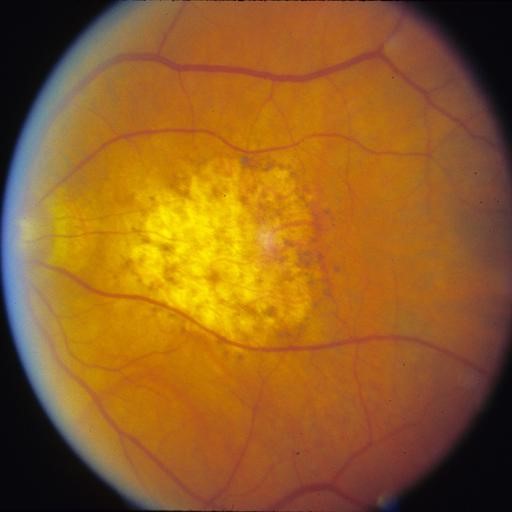
Analyses were carried out on a Linux machine running Ubuntu 18.04, with 32GB of memory and a Titan X 12GB GPU. Preprocessing and model development were carried out with Python 3.7 [33], with Keras 2.2.4 [10] and Tensorflow [1] used for deep learning. Statistical analysis of the results was conducted in R 2.4.4 [24], with the pROC [25], ReportROC [13], dca [28] and PredictABEL [20] packages..

In model development, we used the Adam optimizer [19] with an initial learning rate of 0.0001, which was reduced to one-fifth if performance did not improve after five epochs. Binary cross-entropy was used as the loss function. Early stopping was utilized with a patience of 10 epochs, and the best model was chosen according to the validation Youden’s index [37] to prevent overfitting.



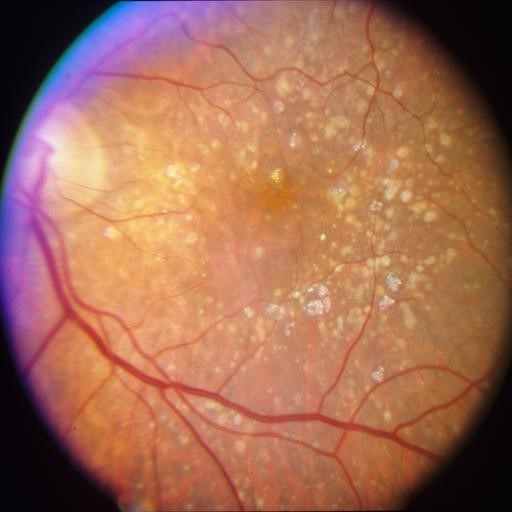
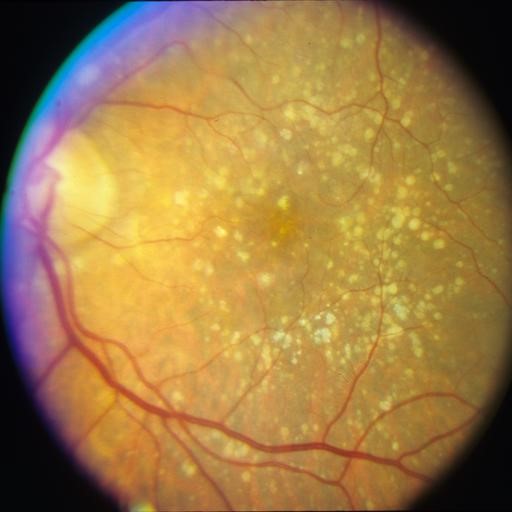
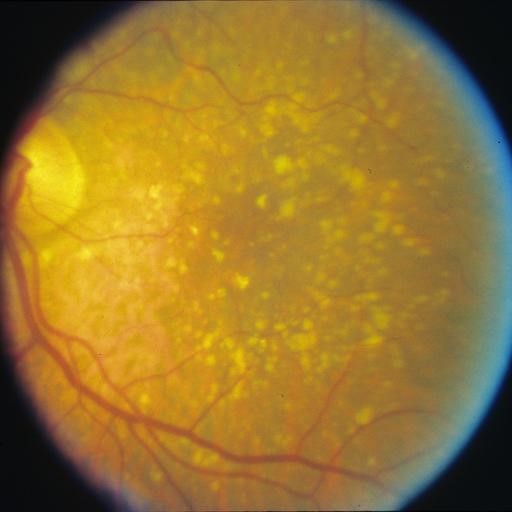
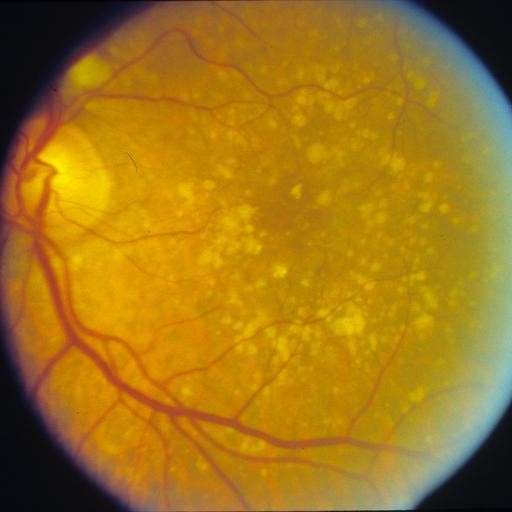
0 years 2.5 years 3 years 4 years

Early/intermediate Early/intermediate Early/intermediate Early/Intermediate (a) Non-progressing

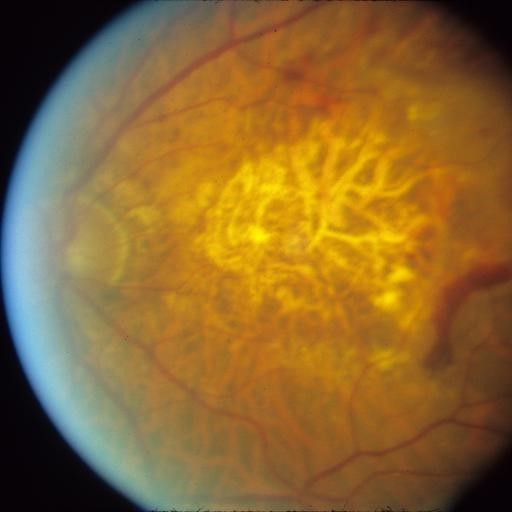
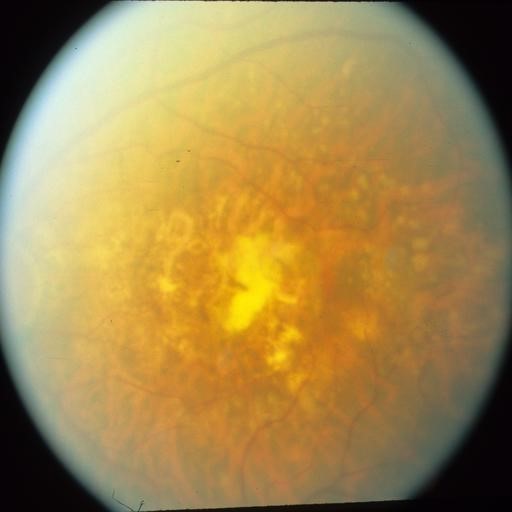
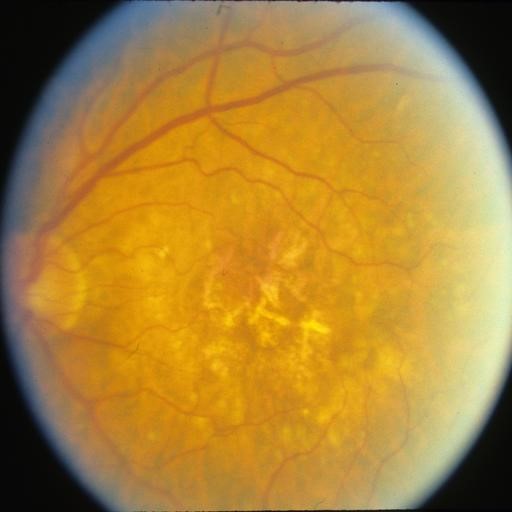
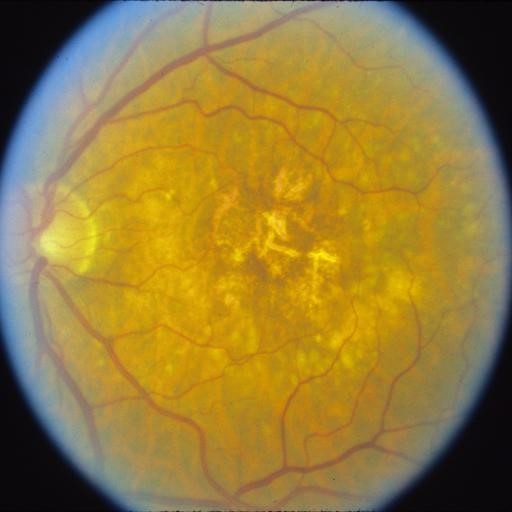


0 years 2 years 3 years 5 years

Early/intermediate Early/intermediate Early/intermediate GA (b) Progressing to GA



|  |  |  |
| --- | --- | --- |
| 0 years | 2 years 4 years | 7 years |
| Early/intermediate | Early/intermediate Early/intermediate  (c) Progressing to nAMD | nAMD |



0 years 0.5 years 3.5 years 6.5 years

Early/intermediate Early/intermediate Early/intermediate GA & nAMD

(d) Progressing to GA and nAMD

**Fig.2.** Sample images of (a) non-progressing patient, (b) patient progressing to geographic atrophy (GA) at time point four, (c) patient progressing to neovascular AMD (nAMD) at time point four, (d) patient progressing to both GA and nAMD at time point four.

**Table 1.** The data were split into training, validation, and testing.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Validation | Testing |
| Progressing | 272 | 91 | 90 |
| Non-progressing | 2670 | 890 | 890 |

## Preprocessing

We first cropped the images to the boundary. An automated algorithm calculated the difference between the image pixel values and the black background; this

difference map was then used to calculate the bounding box, and the image was cropped to this bounding box. All images were rescaled to 256×256 pixels, using the Lanczos resampling [31], to reduce computational requirements. Images of right eyes were flipped horizontally, such that the optic disc is located on the left-hand side of each image. Finally, pixel values were rescaled to between 0 and

1.

## Metrics

We assess model performance in four areas: overall performance (Brier score [8]), discrimination (area under the receiver operating characteristic (AUC) [16], sensitivity, and specificity), reclassification (net reclassification index (NRI) [21]), and clinical usefulness (decision curve analysis (DCA) [34]).

The overall performance of the model is evaluated using the Brier score, which is defined as the mean square difference between the predicted probability and the observed outcome, with lower values indicating better performance. As the data is highly imbalanced, with only around 10% of patients progressing to AMD, we report the area under the receiver operating characteristic (AUC), as well as the optimum sensitivity and specificity, based on Youden’s index [37]. Confidence intervals are reported to show whether differences in results are statistically significant. We used De Long’s method [12] to calculate the 95% confidence intervals and De Long’s test to test for a statistically significant difference between AUCs. For sensitivity and specificity, we calculated bootstrapped 95% confidence intervals with 2000 samples. Youden’s index was used in model selection and is not reported; however, the index can be easily calculated by *J* = *sensitivity* + *specificity* − 1. The NRI is a measure of how well a new model reclassifies subjects, with positive values indicating that the new model classifies subjects better than the previous model. DCA displays graphically how well diagnostic and prognostic models perform by calculating the net benefit at various thresholds and showing at what threshold a model ceases to have a net benefit.

## Results

The proposed method attains a Brier score of 0.050, an AUC of 0.966 (95% CI: 0.947, 0.984), a sensitivity of 0.878 (95% CI: 0.810, 0.945), and a specificity of 0.930 (95% CI: 0.914, 0.947). The single time point without scaling method attained a Brier score of 0.121, an AUC of 0.873 (95% CI: 0.838, 0.907), a sensitivity of 0.911 (95% CI: 0.852, 0.970), and a specificity of 0.764 (95% CI: 0.736, 0.792). The single time point without scaling method attained a Brier score of 0.114, an AUC of 0.881 (95% CI: 0.856, 0.906), a sensitivity of 0.956 (95% CI: 0.913, 0.998), and a specificity of 0.816 (95% CI: 0.790, 0.841). The single time point without scaling method attained a Brier score of 0.124, an AUC of 0.889 (95% CI: 0.862, 0.915), a sensitivity of 0.911 (95% CI: 0.852, 0.970), and a specificity of 0.792 (95% CI: 0.765, 0.819).

The results show that the proposed method achieves a statistically significant improvement in AUC and specificity over the other methods, with a nonsignificant difference in sensitivity. De Long’s test for difference in AUCs gave a p-value of *<*0.0001 against the second-highest AUC, indicating that the improvement in AUC is highly significant. These results are presented in Table 2, with the ROC curve displayed graphically in Figure 3.

Compared to the proposed VAR with the scaling method, the net reclassification improvements were 0.146 (95% CI: 0.053, 0.239) net improvement over the single time point without scaling, 0.334 (95% CI: 0.225, 0.443) over the single time point with scaling, and 0.176 (95% CI: 0.065, 0.287) over the VAR without scaling method, showing a significant improvement in classification over all other models. The decision curve, shown in Figure 4, also shows that the VAR with scaling method provides an increased net benefit over the other methods.

**Table 2.** Performance metrics in the testing set. Brier score. Area Under the Receiver Operating Characteristic Characteristic (AUC) with 95% confidence intervals constructed by De Long’s method. Sensitivity and specificity with 95% confidence intervals constructed by bootstrapping with 2000 samples. The proposed method achieves an improved Brier score, statistically significantly higher AUC and sensitivity, and a non-significant difference in sensitivity. Bold indicates statistically significant results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Brier | AUC | Sensitivity | Specificity |
| Single time point without scaling | 0.121 | 0.873  (0.838, 0.907) | 0.911  (0.852, 0.970) | 0.764  (0.736, 0.792) |
| Single time point with scaling | 0.114 | 0.881  (0.856, 0.906) | 0.956  (0.913, 0.998) | 0.816  (0.790, 0.841) |
| VAR  without scaling | 0.124 | 0.889  (0.862, 0.915) | 0.911  (0.852, 0.970) | 0.792  (0.765, 0.819) |
| VAR  with scaling | 0.050 | **0.966**  **(0.947, 0.984)** | 0.878  (0.810, 0.945) | **0.930**  **(0.914, 0.947)** |

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

Specificity

Sensitivity

Method

Single time point without scaling

Single time point with scaling

VAR without scaling

VAR with scaling

**Fig.3.** Receiver operating characteristic curve of the proposed model and comparison models.



**Fig.4.** Decision curve showing the net benefit of the treat all approach, the single time point, and the proposed method. Models with a higher threshold before hitting zero net benefit are considered to more clinically useful. This shows graphically that our model has a greater clinical usefulness over the single time point method.

## Class activation maps

Class activation maps [38] visualize how the network reaches the decision. Class activation maps showed that areas with high levels of drusen are considered important, which is expected. The class activation maps suggest that the model

is correctly identifying features in the majority of images and is working as intended. In earlier images, the class activation maps are not as accurate as those of later images; this suggests that the model becomes more confident of the outcome as the disease progresses, which is also true of human graders. Examples of class activation maps are shown in Figure 5; red shows areas of high importance and blue shows areas of low importance.

# Conclusions

We have proposed a novel method to predict the future course of diseases, utilizing multiple longitudinal images with uneven time intervals between images. The proposed method allows implicit feature extraction to be trained simultaneously alongside a traditional statistical model, namely an autoregressive model. Our method extracts feature vectors from previous images and aims to predict the feature vector of some future disease using an autoregressive model, which is then classified.

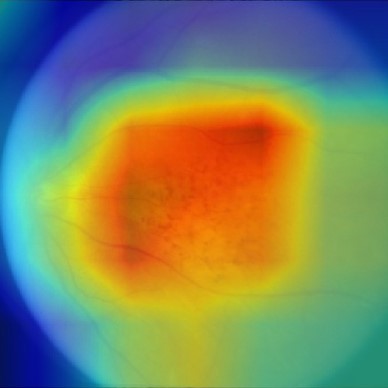
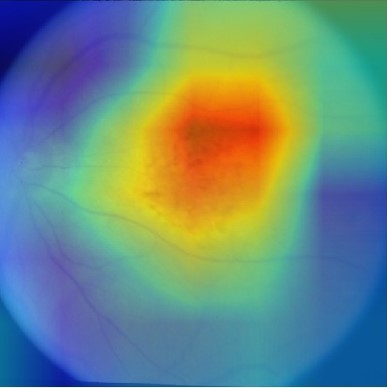
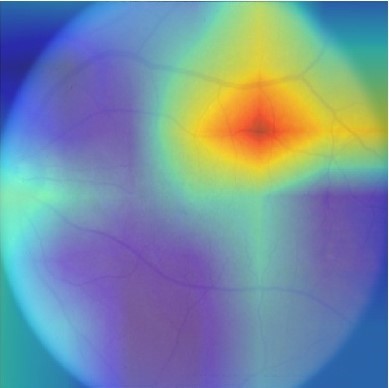
Experiments on a dataset of age-related macular degeneration showed good performance, with particularly high specificity, which is desirable in a screening setting. Class activation maps display that the algorithm is identifying the expected features of an image as being relevant to the final prediction. The method is easily generalizable to many diseases and modalities and requires only the disease outcome as ground truth during training. Future models may include clinical and genetic variables to improve the prediction further.

# References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg,

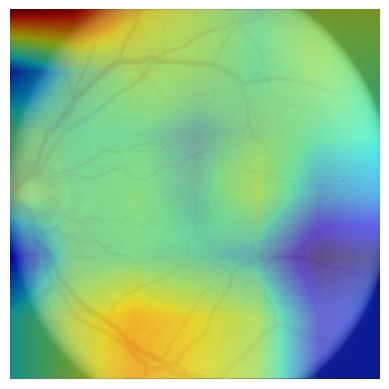
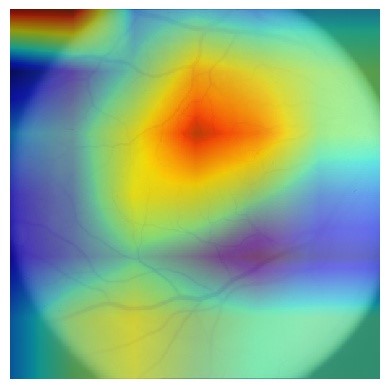
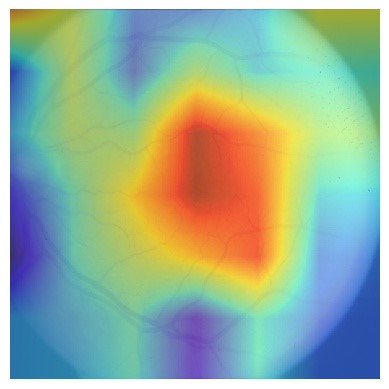
J., Man´e, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Vi´egas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng., X.: Tensorflow: Large-scale machine learning on heterogeneous systems (2015), Software available from tensorflow.org

1. Age-Related Eye Disease Study Research Group: The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. Controlled clinical trials **20**(6), 573–600 (1999). https://doi.org/10.1016/s0197-2456(99)00031-8
2. Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., Prunotto, M.: Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Digit Med **2**, 92 (2019). https://doi.org/10.1038/s41746-019-0172-3
3. Arenja, N., Andre, F., Riffel, J.H., Siepen, F.a.d., Hegenbart, U., Scho¨nland, S., Kristen, A.V., Katus, H.A., Buss, S.J.: Prognostic value of novel imaging parameters derived from standard cardiovascular magnetic resonance in high risk patients with systemic light chain amyloidosis. Journal of Cardiovascular Magnetic Resonance **21**(1), 53 (2019). https://doi.org/10.1186/s12968-019-0564-1



0 years 2 years 3 years

(a) Correctly classified as progressing.



0 years 2 years 3 years

(b) Failed example. Incorrectly classified as non-progressing.

**Fig.5.** Two example class activation maps from the testing dataset. Areas in red are considered more important by the network during classification. (a) The centre of the image show a high concentration of drusen, which is expected to be important. At 0 years, the algorithm highlights an incorrect area, as drusen is not easily observed at this time point. At 2 years, drusen is more visible and the algorithm begins to correctly identify the expected features. At 3 years, drusen is highly visible and the algorithm correctly identifies the area of interest. (b) A failed example where the algorithm incorrectly classified the image as non-progressing. The class activation map shows no particular region of interest in the final image. The disease is difficult to see and the second image appears to be from a different eye to the other two, suggesting a mistake within the dataset, which confuses the network and leads to an incorrect prognosis.

1. Babenko, B., Balasubramanian, S., Blumer, K.E., Corrado, G.S., Peng, L., Webster, D.R., Hammel, N., Varadarajan, A.V.: Predicting progression of age-related macular degeneration from fundus images using deep learning. arXiv preprint arXiv:1904.05478 (2019)
2. Banerjee, I., de Sisternes, L., Hallak, J., Leng, T., Osborne, A., Durbin, M., Rubin, D.: A deep-learning approach for prognosis of age-related macular degeneration disease using SD-OCT imaging biomarkers. arXiv preprint arXiv:1902.10700 (2019)
3. Bridge, J., Harding, S., Zheng, Y.: Development and validation of a novel prognostic model for predicting amd progression using longitudinal fundus images. BMJ

Open Ophthalmology **5**(1) (2020). https://doi.org/10.1136/bmjophth-2020-000569

1. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1–3 (1950)
2. Choi, H., Jin, K.H.: Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behavioural Brain Research **344**, 103–109

(2018). https://doi.org/10.1016/j.bbr.2018.02.017

1. Chollet, F., et al.: Keras (2015), https://keras.io
2. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**(2), 187–220 (1972), www.jstor.org/stable/2985181
3. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988)
4. Du, Z., Hao, Y.: reportROC: An Easy Way to Report ROC Analysis (2019), https://CRAN.R-project.org/package=reportROC, R package version 3.4
5. Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M.E., Linkohr, B., Peters, A., Heid, I.M., Palm, C., Weber, B.H.F.: A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology **125**(9), 1410–1420 (2018). https://doi.org/10.1016/j.ophtha.2018.02.037
6. Gregor, K., Danihelka, I., Mnih, A., Blundell, C., Wierstra, D.: Deep autoregressive networks. arXiv preprint arXiv:1310.8499 (2013)
7. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Jama **247**(18), 2543–2546 (1982), https://jamanetwork.com/journals/jama/article-abstract/372568
8. Hilario, A., Sepulveda, J.M., Perez-Nun˜ez, A., Salvador, E., Millan, J.M., Hernandez-Lain, A., Rodriguez-Gonzalez, V., Lagares, A., Ramos, A.: A prognostic model based on preoperative MRI predicts overall survival in patients with diffuse gliomas. American Journal of Neuroradiology **35**(6), 1096 (2014). https://doi.org/10.3174/ajnr.A3837
9. Keane, P.A., Patel, P.J., Liakopoulos, S., Heussen, F.M., Sadda, S.R., Tufail, A.: Evaluation of age-related macular degeneration with optical coherence tomography. Survey of Ophthalmology **57**(5), 389–414 (2012).

https://doi.org/10.1016/j.survophthal.2012.01.006

1. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
2. Kundu, S., Aulchenko, Y.S., Janssens, A.C.J.: PredictABEL: Assessment of Risk Prediction Models (2014), https://CRAN.R-project.org/package=PredictABEL, R package version 1.2-2
3. Leening, M.J., Vedder, M.M., Witteman, J.C., Pencina, M.J., Steyerberg, E.W.: Net reclassification improvement: Computation, interpretation, and controversies:

A literature review and clinician’s guide. Annals of Internal Medicine **160**(2), 122– 131 (2014). https://doi.org/10.7326/m13-1522

1. Owen, C.G., Jarrar, Z., Wormald, R., Cook, D.G., Fletcher, A.E., Rudnicka, A.R.: The estimated prevalence and incidence of late stage age related macular degeneration in the uk. British Journal of Ophthalmology **96**(5), 752 (2012). https://doi.org/10.1136/bjophthalmol-2011-301109
2. Priestley, M.: Non-linear and Non-stationary Time Series Analysis. Academic Press (1988)
3. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2019), https://www.R-project.org/
4. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Mu¨ller, M.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **12**(1), 77 (2011). https://doi.org/10.1186/1471-210512-77
5. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
6. de Sisternes, L., Simon, N., Tibshirani, R., Leng, T., Rubin, D.L.: Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression. Invest Ophthalmol Vis Sci **55**(11), 7093–103 (2014). https://doi.org/10.1167/iovs.14-14918
7. Sjoberg, D.D.: dca: Decision Curve Analysis (2020), R package version 0.1.0.9000
8. Steyerberg, E.W., Moons, K.G.M., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H., Altman, D.G., Group, P.: Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS medicine **10**(2), e1001381–e1001381 (2013)
9. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826
10. Turkowski, K.: Filters for common resampling tasks (1990)
11. Uria, B., Coˆt´e, M.A., Gregor, K., Murray, I., Larochelle, H.: Neural autoregressive distribution estimation. The Journal of Machine Learning Research **17**(1), 7184– 7220 (2016)
12. Van Rossum, G., Drake Jr, F.L.: Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam (1995)
13. Vickers, A.J., Elkin, E.B.: Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making **26**(6), 565–574 (2006). https://doi.org/10.1177/0272989x06295361
14. Williams, B.M., Burgess, P.I., Zheng, Y.: Chapter 13 - Drusen and macular degeneration, pp. 245–272. Academic Press (2019). https://doi.org/10.1016/B978-0-08102816-2.00013-7
15. Yan, Q., Weeks, D.E., Xin, H., Huang, H., Swaroop, A., Chew, E.Y., Ding, Y., Chen, W.: Deep-learning-based prediction of late age-related macular degeneration progression. medRxiv p. 19006171 (2019). https://doi.org/10.1101/19006171
16. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
17. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition. pp. 2921–2929