BioMedInformatics

MDPI

Article

Machine Learning for Extraction of Image Features Associated with Progression of Geographic Atrophy

Janan Arslan [1,2] and Kurt Benke [3,4,*]

[1] Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013 Paris, France; janan.arslan@icm-institute.org
[2] Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye & Ear Hospital, East Melbourne, VIC 3002, Australia
[3] School of Engineering, University of Melbourne, Parkville, VIC 3010, Australia
[4] Graeme Clark Institute for Biomedical Engineering, Parkville, VIC 3010, Australia
* Correspondence: kbenke@unimelb.edu.au

Abstract: Background: Several studies have investigated various features and models in order to understand the growth and progression of the ocular disease geographic atrophy (GA). Commonly assessed features include age, sex, smoking, alcohol consumption, sedentary lifestyle, hypertension, and diabetes. There have been inconsistencies regarding which features correlate with GA progression. Chief amongst these inconsistencies is whether the investigated features are readily available for analysis across various ophthalmic institutions. Methods:In this study, we focused our attention on the association of fundus autofluorescence (FAF) imaging features and GA progression. Our method included feature extraction using radiomic processes and feature ranking by machine learning incorporating the algorithm XGBoost to determine the best-ranked features. This led to the development of an image-based linear mixed-effects model, which was designed to account for slope change based on within-subject variability and inter-eye correlation. Metrics used to assess the linear mixed-effects model included marginal and conditional R^2, Pearson's correlation coefficient (r), root mean square error (RMSE), mean error (ME), mean absolute error (MAE), mean absolute deviation (MAD), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and loglikelihood. Results: We developed a linear mixed-effects model with 15 image-based features. The model results were as follows: $R^2 = 0.96$, $r = 0.981$, RMSE = 1.32, ME = -7.3×10^{-15}, MAE = 0.94, MAD = 0.999, AIC = 2084.93, BIC = 2169.97, and log likelihood = −1022.46. Conclusions: The advantage of our method is that it relies on the inherent properties of the image itself, rather than the availability of clinical or demographic data. Thus, the image features discovered in this study are universally and readily available across the board.

Keywords: age-related macular degeneration; geographic atrophy; fundus autofluorescence imaging; statistical modelling; feature ranking; mixed-effects modelling; disease progression

check for updates

Citation: Arslan, J.; Benke, K. Machine Learning for Extraction of Image Features Associated with Progression of Geographic Atrophy. *BioMedInformatics* **2024**, *4*, 1638–1671. https://doi.org/10.3390/biomedinformatics4030089

Academic Editors: M. Michael Gromiha and Alexandre G. De Brevern

Received: 27 February 2024
Revised: 26 March 2024
Accepted: 5 June 2024
Published: 2 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Several studies to date have investigated various features and models to understand the growth and progression of geographic atrophy (GA) in age-related macular degeneration. Some commonly assessed clinical and demographic features have included age, sex, smoking, alcohol consumption, and sedentary lifestyle, with outcomes suggesting mixed results with respect to the association of these variables with GA. Other features have been similarly explored, including hypertension, diabetes, hyperlipidaemia, and body mass index, although these explorations have not generally been fruitful [1–4]. These features have been used in several different models to explain both the progression of the disease and the variability in progression between patients.

Pfau et al. (2019) used shape-descriptive features in a linear mixed-effects model with a two-level random effect of eye and patient [5]. Liefers et al. (2020) also included many

shape-based feature associations in their linear regression model [6]. For GA spatial prediction, Niu et al. (2016) used the random forest and 19 quantitative image features which characterised the shape and size of GA, as well as its axial location (e.g., area, maximum radius, eccentricity, etc.) [7]. Pfau et al. (2020) also assessed spatial prediction using a mixed-effects logistic model that included features such as retinal pigment epithelium (RPE)-based diagnostic groups, pixel-wise locations, and follow-up time. Schmidt-Erfurth et al. (2020) used a linear regression model to determine an association between GA progression and HRF [7–9]. For prediction of GA visual function, Künzel et al. (2020) used the linear regression model along with more clinical features, such as best-corrected VA, while Pfau et al. (2020) used the random forest approach along with image features and patient reliability indices to estimate retinal sensitivity in GA progression [10,11]. These publications used a combination of metrics in post-estimation evaluations, including the mean absolute error (MAE), the mean square error (MSE), and the coefficient of determination (R^2), in evaluating the overall fit of their respective models. These papers used various modelling and machine learning (ML) techniques in estimating GA progression; however, none of these techniques was justified based on clinical assumptions or demonstrated the fundamental reasoning behind their choice of modelling.

For this study, the linear mixed-effects model was used for several reasons. Firstly, in our previous publications, the linear progression model was shown to be most representative of GA progression using *real clinical presentations*. In our previous work, we demonstrated that the sigmoidal growth curve may best represent GA progression from start to finish. However, the research was based on a limited number of case studies of patients with the highest number of clinical presentations, where the aim was to more accurately estimate the lower and upper tail ends of the sigmoidal growth curve [12,13]. Future iterations of this work will involve the collection of additional, real-time clinical data to explain those lower and upper tail ends. Such data will include nascent GA. However, the focus of this study was on creating a prediction model based on the real clinical presentations and thus narrowed on the peak growth rates of GA progression (i.e., the linear portion of the sigmoidal growth curve). Furthermore, given that peak growth occurs in the linear portion of the sigmoidal growth curve, there were practical reasons for focusing simply on the linear portion.

The reason for the choice of the mixed-effects model was two-fold: firstly, patient-level variation needs to be accounted for given that there is high variability in terms of GA progression from patient to patient; secondly, as discussed by Arslan et al. (2021), inter-eye correlation may affect the results if treated as a single probability distribution [13]. Ying et al. (2017) outlined potential avenues in which inter-eye correlation could be tackled while highlighting their advantages and disadvantages [14]. The most promising of these methods is the use of the mixed-effects model to ensure that the regression coefficients remain unbiased and that the variance estimators are correct and thus that the model average is not affected by correlations between the left and right eye.

An additional motive for the selection of the mixed-effects model was that it is suitable for time-series and hierarchical datasets given that it is a multilevel model. Biological and medical data are complex, and care needs to be taken to avoid spurious or inflated associations. There are several possible causes of confounding, including population structure (the existence of major subgroups in a population), cryptic relatedness (the existence of small groups of related individuals), and environmental factors (environmental differences between subpopulations or geographic locations) [15–18]. Several methods have been suggested to control for these confounders, one of which includes mixed-effects modelling, where a set of random effects is fitted for each individual [18]. Mixed-effects models are well suited for analysis of biological/medical data [19] and are flexible and powerful statistical models for controlling stratification, relatedness, and confounding factors [20–22].

We have seen in the literature some models utilising image-based features, with particular focus on shape features, as highlighted in the preceding paragraph. The current study

also utilised image features, specifically FAF-based features, in modelling GA progression. The analysis applied a process referred to as *radiomics*—a quantitative feature extraction approach specifically designed for the extraction of data and metrics from medical and clinical images [23]. However, the radiomic process can lead to the extraction of large numbers of features. This can make the process of univariate feature selection incredibly strenuous when traditional statistical techniques are used. Furthermore, while we tried to identify the most suitable features for the model, we also wished to select the features that were ranked in order of highest importance to lowest importance to determine the features most relevant to the disease of interest. To do this using traditional statistical techniques would involve the assignment of weights to each variable using the Relative Importance of Variables (RIV) method [24].

The RIV method ranks predictor variables by weights: predictors with larger weights (i.e., close to 1) are considered the most important features, while those with the smallest weights (i.e., close to 0) are considered to be the least important [25]. The null hypothesis for an RIV analysis is that the parameter estimates for all variables are identical and have the same level of importance. To estimate the RIV of variable x_j, the sum of all Akaike weights is required (i.e., the AIC) across all models in the set in which j occurs; the sum of $w_+(j)$ reflects the importance of the variable. This sum is denoted as a numerical value between 0 and 1. The larger the sum of $w_+(j)$ (i.e., the closer it is to 1), the more important the variable is relative to the other variables tested. Using $w_+(j)$, all the variables can be ranked in order of their importance. To obtain these weights, the effect size of each variable needs to be based on model-averaged estimates. In other words, to ensure an accurate reflection of the importance of one variable versus another, a combination of models is required which contains all prospective variables in equal proportions across all models, allowing each variable to be tested on an equal footing. Otherwise, if one variable were to be found more frequently across the test models as compared to the others, this may inadvertently give the more frequently occurring variable the advantage and cause it to rank higher than the others. The process of looking at all potential combinations of variables, creating a new model for each combination of variables, and then assessing each model iteratively while modifying the rank of each variable is laborious and time consuming. For example, from past experience, the process of looking at a combination of 19 variables using the RIV method could take up to several days.

In this paper, over 100 variables were considered, and combinatorial assessment with the RIV method would mean that a machine would need to be on for several days without a break in order to rank the variables in order of importance. This time-intensive process can now be replaced with efficient and expedient ML processes. Thus, in this paper, *feature extraction* was conducted using radiomics, while *feature selection* and weighting was conducted using the ML approach XGBoost. The XGBoost algorithm is a gradient-descent learning method which uses a decision tree in order to rank features in terms of their importance and relevance to the outcome. This method is computationally faster and has several advantages, including the removal of redundant features. This removal of redundant features is an additional bonus, as decision trees are typically immune to multicollinearity, meaning that no further analyses need to be performed to manually determine collinear features during model development. When a decision tree is faced with, say, three variables that are highly correlated and collinear, it will split between the three and identify the one with the greatest performance, thus eliminating the other correlated variables.

In the remainder of this paper, we discuss in greater detail feature extraction, feature selection, linear mixed-effects models, and the most suitable model based on model diagnostics, such as Q-Q plotting, which assesses the assumptions of normality. While this paper presents the most suitable model, it is important to note that this suitability is based on the current available dataset, which predominantly includes image-based features. Image-based features are but one component of a larger picture of a complex disease, and thus the model presented in this paper should be considered a preliminary fit until

additional data become available. This FAF-based linear mixed-effects model could be improved upon in the future with new data, which could include features extracted from other imaging modalities, such as SD-OCT, along with clinical information, such as genetics. The latter would create a more personalised prediction model.

2. Materials and Methods

2.1. Study Design

A retrospective analysis was conducted as a case study using anonymised data from patients who attended the retina clinics at the Royal Victorian Eye and Ear Hospital (RVEEH). The study was approved by the Human Research Ethics Committee of the RVEEH. The study was conducted in accordance with the International Conference on Harmonisation Guidelines for Good Clinical Practice and the tenets of the Declaration of Helsinki. Ethical approval was provided by the Human Research Ethics Committee (HREC: project no. 95/283H/15) of the RVEEH. Written informed consent was obtained from all participants.

2.2. Feature Extraction

The objective of the radiomic process is to obtain large amounts of data extracted and quantified from medical images. Radiomic features were extracted using pyradiomics (v3.0: https://pyradiomics.readthedocs.io/en/latest/, accessed on 19 March 2024), which is a Python package supported by the US National Cancer Institute [26]. There is a total of 107 (Table 1) relevant radiomic features for FAF images (all 2D features) [26]. Examples of these procedures include the following:

- *Shape-based (2D) features*, which include size, shape, volume, and surface measurements.
- *First-order statistics*, which describe intensities and distributions.
- *Neighbouring Gray Tone Difference Matrix (NGTDM) features*, which quantify average gray levels from neighbouring gray levels.
- *Gray-Level Dependence Matrix (GLDM) features*, which look at gray-level dependencies in an image.
- *Gray-Level Run-Length Matrix (GLRLM) features*, which quantify gray-level runs, such as consecutive regions of gray levels.
- *Gray-Level Co-occurrence Matrix (GLCM) features*, which are second-order joint probability functions of an image.
- *Gray-Level Size Zone Matrix (GLSZM) features*, which quantify gray-level zones.

Table 1. Radiomic and clustering features extracted from FAF images. More information regarding each radiomic feature can be found at https://pyradiomics.readthedocs.io/en/latest/features.html (accessed on 12 August 2021). Clustering features were the result of our previous work [27].

Feature Category	Features
Shape (2D)	Maximum 2D Diameter Row, Volume, Elongation, Flatness, Maximum 3D Diameter, Least Axis, Surface Volume Ratio, Major Axis, Surface Area, Maximum 2D Diameter Slice, Minor Axis, Maximum 2D Diameter Column, Sphericity
Neighbouring Gray Tone Difference Matrix (NGTDM) Features	Busyness, Coarseness, Complexity, Strength, Contrast
First-Order Statistics	10th Percentile, Variance, Interquartile Range, Skewness, Mean, Energy, Uniformity, Root Mean Squared, 90th Percentile, Kurtosis, Mean Absolute Deviation, Range, Total Energy, Maximum, Median, Robust Mean Absolute Deviation, Minimum, Standard Deviation, Entropy
Gray-Level Dependence Matrix (GLDM) Features	Gray-Level Non-Uniformity, Dependence Entropy, Small Dependence Low Gray-Level Emphasis, Gray-Level Variance, Dependence Non-Uniformity Normalised, Large Dependence High Gray-Level Emphasis, Large Dependence Emphasis, Large Dependence Low Gray-Level Emphasis, Small Dependence High Gray-Level Emphasis, Dependence Variance, High Gray-Level Emphasis, Dependence Non-Uniformity, Low Gray-Level Emphasis, Small Dependence Emphasis

Table 1. *Cont.*

Feature Category	Features
Gray-Level Run-Length Matrix (GLRLM) Features	Gray-Level Non-Uniformity, Run Percentage, Run-Length Non-Uniformity Normalised, Short Run High Gray-Level Emphasis, Long Run Emphasis, Short Run Low Gray-Level Emphasis, Gray-Level Variance, Run-Length Non-Uniformity, Short Run Emphasis, Run Variance, Run Entropy, Gray-Level Non-Uniformity Normalised, Low Gray-Level Run Emphasis, Long Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, High Gray-Level Run Emphasis
Gray-Level Co-occurrence Matrix (GLCM) Features	Imc1, Sum Average, Correlation, Idn, Joint Entropy, Difference Entropy, Cluster Tendency, Joint Energy, Difference Variance, Id, Joint Average, Idm, Cluster Shade, Imc2, Inverse Variance, Cluster Prominence, Sum Squares, Sum Entropy, Difference Average, Contrast, Idmn, Maximum Probability, Autocorrelation
Gray-Level Size Zone Matrix (GLSZM) Features	Gray-Level Non-Uniformity, Large Area High Gray-Level Emphasis, High Gray-Level Zone Emphasis, Size Zone Non-Uniformity, Small Area High Gray-Level Emphasis, Zone Variance, Gray-Level Variance, Large Area Low Gray-Level Emphasis, Size Zone Non-Uniformity Normalised, Gray-Level Non-Uniformity Normalised, Small Area Emphasis, Large Area Emphasis, Small Area Low Gray-Level Emphasis, Low Gray-Level Zone Emphasis, Zone Percentage, Zone Entropy
Clustering Features	Early-Stage Hyperfluorescence Cluster, Late-Stage Hyperfluorescence Cluster, Lesion Cluster

Radiomic features were extracted for all GA regions of interest, including lesions and early-stage, intermediate-stage, and late-stage hyperfluorescence areas. To extract these features, the binarised masks created using the techniques described by Arslan et al. (2021; 2023) were utilised [27,28]. These binarised masks were coupled with their original FAF images. When radiomic features were being extracted, first, the original image was inputted and then the binarised masks highlighted the areas within the original image from which information needed to be extracted. See Figure 1 for an illustration of this process.

In addition to the radiomic features, clustering results from Arslan et al. (2023) were also included as additional variables [27]. Specifically, the variables named "early_cluster" (i.e., early-stage hyperfluorescence clustering), "late_cluster" (late-stage hyperfluorescence clustering), and "lesion_cluster" (lesion clustering) were extracted and added to the database. Intermediate-stage clustering was not included as an additional variable given that distinguishable clustering groups were not determined for this GA feature. This created a total of 431 (i.e., 4 GA areas × 107 radiomic features + 3 cluster variables) features to evaluate.

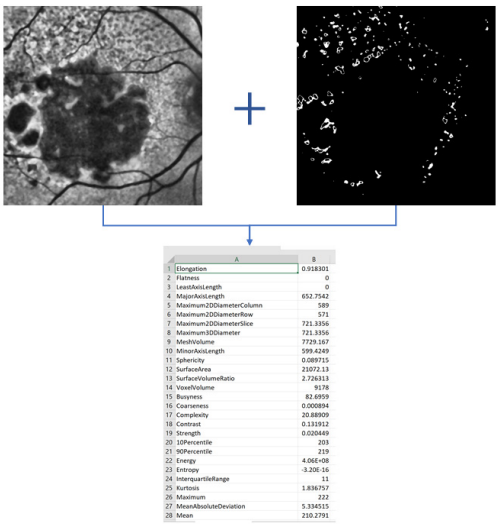


Figure 1. Process of radiomic feature extraction. The original image was coupled with the binarised output. This mask indicated the regions within the original image from which features should be extracted. In the above example, features were being extracted for the late-stage hyperfluorescence regions of GA. GA: Geographic atrophy.

2.3. Feature Selection

Feature selection is the process of ranking the variables in terms of importance relative to the outcome of interest (i.e., the total area of GA in mm^2/year). Feature selection was conducted using an ML approach and Python (v3.6), namely, XGBoost (v1.4.0: https://xgboost.readthedocs.io/en/latest/index.html, accessed on 19 March 2024). XGBoost (eXtreme Gradient Boosting) is a gradient-descent learning method which uses a decision tree in order to rank features in terms of their importance and relevance to an outcome. XGBoost works by identifying the most appropriate "split points". It calculates the importance of these split points relative to the outcome measure and calculates overall how well these split points perform across different iterations. In some ways, this process is similar to the RIV method previously highlighted in the Introduction section, although XGBoost has several advantages. Its greatest advantage is that it can combine many weak learners and combine the weak learners into strong learners. Weak learners refer to a single decision tree which, if used by itself, would be only marginally better than randomly guessing an outcome. For example, if we were to look at which factors contribute to the successful graduation of high-school students and, amongst the many potential factors involved, decided to solely focus on familial factors (e.g., the age and assistance of parents), then we would have a subpar model that does not comprehensively cover all the potential features that could contribute to the success of students. Therefore, any model based on these weak learners would be no better than guessing how well a student is going to perform. However, if we were to combine many weak learners (e.g., the hours a student spends studying, the age and assistance of their parents, the financial stability of the family, tutoring that the student might receive, the child's work ethic, extracurricular or volunteer work outside of school hours, etc.), then we would have more information that we could assess collectively and use to create better and more comprehensive suggestions as to which of these features is best suited to the outcome of interest. The plethora of information with strong learners leads to better success rates than with individual weak learners [29].

In addition to the combination of weak learners to build strong learners, XGBoost is also capable of building decision trees in parallel rather than sequentially, as other decision trees typically do. This parallel computing format in assessing the problem leads to the efficiency and speed for which XGBoost is known and has been popularised. Furthermore, XGBoost has the ability to leverage residual patterns that help in creating prediction errors which are ten times lower than those of other decision trees, such as the random forest. This leverage of residuals, along with its ability to control model complexity and reduce overfitting via built-in regularisation, contributes to its predictive performance. The added benefit of using XGBoost is its ability to remove redundant features automatically, and thus later in the modelling process there is no need to conduct multicollinearity assessment. XGBoost is also capable of handling missing values with ease [29].

Despite several advantages, XGBoost is still prone to variability and uncertainty, as highlighted in [12,13], because it is a data-dependent method. Therefore, for completeness, the XGBoost method was run five times (i.e., five-fold cross-validation). The five-fold cross-validation method was used here to ensure that the top selected features were consistently selected. Additionally, to ensure that the topmost important and relevant features were the focus, XGBoost was programmed to select and rank the top 30 of the 431 image features. The objective of selecting the top 30 was to ensure that the selected model was a parsimonious model.

2.4. Mixed-Effects Model

The development of the linear mixed-effects model and all subsequent statistical analyses were performed using the statistical software R version 4.0.0 [30]. Mixed-effects models are used to describe relationships between response and predictor variables in data that are grouped based on one or more classifications [31]. Mixed-effects models explicitly specify the mean and covariance structure, incorporating two types of parameters: fixed and random effects [32,33]. Fixed effects refer to predictors that affect a response variable.

Random effects, however, refer to effects on a response variable generated by variation within and among levels of a predictor variable [32]. Population structure is the fixed effect in a mixed-effects model, while relatedness among individuals is incorporated as a variance–covariance structure of the random effect [34]. Mixed-effects models have gained considerable popularity and are considered useful in the analysis of longitudinal data, modelling of complex clustered data, penalised log likelihood, etc. [19,35].

There are advantages to using mixed models in medical applications. Medical data are often clustered (e.g., a study may be recorded at several centres, hospitals, etc.). The design of medical studies can be described as hierarchical, and wider inferences can be made by fitting the clustering effect as random. Repeated measurements are also common in medicine-based studies, and it is not uncommon for several observations to be missing. The advantage of using mixed-effects models in such studies is that allowances can be made for missing data and the hierarchical clustering [36].

The data for this study were longitudinal and consisted of repeated observations by individual subjects over time. The interest lies in the effects that are common and different among all individuals [37]. The mixed-effects model allows the capture of among-subject variations. In this thesis, the linear mixed-effects model was used as (1) it was confirmed that the clinical presentations can be represented by the linear regression model and (2) it is a useful model that can account for inter-eye correlation, as highlighted in [13]. Linear mixed-effects models are an extension of regular linear models. Traditional linear models only use a single random term, the residual error. Linear mixed-effects models allow the specification of more than one random term—a useful feature, as it is more accurate to think of an effect coming from a specific normal distribution rather than a fixed value [38].

With N independent sampling units (i.e., patients), the linear mixed-effects model for the ith person may be written as follows:

$$Y_i = X_i\beta_i + Z_i u_i + \varepsilon_i \quad (1)$$

where Y_i represents the response variable for the ith person, X_i is an $n_i \times p$ design matrix for the p-vector of fixed effect β, and Z_i is an $n_i \times q$ design matrix associated with the q-vector of random effects u_i that represents subject-specific regression coefficients. The error term, ε_i, is assumed to be normally distributed with mean zero and to be independent of the random effects [39]. The use of linear mixed-effects models counters the multiple drawbacks that are normally associated with traditional random-effects modelling, such as the following [40]:

- Deficiencies in statistical power with the use of repeated observations;
- The lack of adaptability due to missing data;
- Disparate methods for treating continuous and categorical responses;
- Dubious methods for modelling heteroscedasticity and non-spherical error variance.

There were multiple measurements for each subject (i.e., patient) in the data; thus, we needed to incorporate random effects into the model to account for the variation in outcomes. To account for within-subject dependencies, a subject-specific latent variable (i.e., a random effect) had to be included in the model. Typically, an additional random effect is included for each regression coefficient that is expected to vary among subjects. For example, in dose–response settings, one may account for baseline heterogeneity through a random intercept and for heterogeneity in susceptibility through a random slope, with these two factors being potentially correlated [41]. To account for this heterogeneity, the random effects used across all our tested models included *subjects* (i.e., an ID variable) and *eyes* (i.e., the left and right eye). These are represented in the analysis as (*subject* | *eye*). The use of the random effect *eye* accounts for the random intercept. The random effect *subject* accounts for the random slope. The lme4 (v1.1-27) package created by Bates and Maechler and maintained by Ben Bolker was used in the creation and analysis of the linear mixed-effects models [42]. Specifically, for the linear mixed-effects model, the lmer() function was used.

2.5. Measure of Outcome

Naturally, the main focus is on the total area of growth in mm^2/year (i.e., Y_i). In [13], the popular square-root transformation was also tested, and it was confirmed that the transformation did not impact the residuals of the linear model. However, for completeness, the transformation was again assessed in the presence of image features to see if it had an impact on the predictive power of the linear mixed-effects model and/or its model diagnostics, in contrast to including residuals. In addition to the square-root transformation, another rarely assessed transformation was also investigated—the log transformation. The log transformation has been claimed to be a transformation better suited for use in GA modelling. Brader et al. (2015) investigated the progression rate and pattern of incident GA. Using subjects who developed GA in untreated eyes in the *Complications of Age-Related Macular Degeneration Prevention Trial* (CAPT; ClinicalTrials.gov (10 June 2021) identifier: NCT00000167), their findings contradicted previous studies that used the square-root transformation by suggesting that the log transformation is more useful in the assessment of GA progression and that the square-root transformation failed to eliminate the dependence on initial size. Their log-transformed growth model, which they suggest best fits their data (not available), yielded an enlargement rate of 56% per year [43].

While the log transformation appears to be a suitable surrogate candidate in place of the commonly used square-root transformation, the standard $\log(Y_i)$ transformation would be insufficient in the context of GA. Sometimes, with small GA lesion areas, the log transformation will produce a negative rather than a positive output, which is not suitable given that the model is supposed to predict GA growth over time. Perhaps this may provide an explanation as to why there are not that many studies assessing the $\log(Y_i)$ transformation. In place of the more commonly used log transformation, in this study, the $\log(Y_i + 1)$ transformation was used instead. This transformation is suitable for right-skewed data, such as GA total areas, and the inclusion of the additive constant "1" ensures that the outcome will always be positive and never negative. In this study, three linear mixed-effects models with three outcome measures were tested: the original scale, the square-root-transformed scale, and the $\log(Y_i + 1)$ scale.

2.6. Model Diagnostics and Selection

Information-based criteria, such as the AIC and the *Schwartz* or *Bayesian Information Criterion* (BIC) were used in the model selection process. Although a plethora of information criteria are available for model comparisons, all other criteria are modifications or generalisations of the AIC/BIC [44]. The AIC and BIC criteria are defined as follows [45]:

$$AIC = 2\left[\updownarrow(\hat{\theta}_2) - \updownarrow(\hat{\theta}_1)\right] - 2(p_2 - p_1) \quad (2)$$

$$BIC = 2\left[\updownarrow(\hat{\theta}_2) - \updownarrow(\hat{\theta}_1)\right] - \log_n(p_2 - p_1) \quad (3)$$

where $2\left[\updownarrow(\hat{\theta}_2) - \updownarrow(\hat{\theta}_1)\right]$ is the likelihood ratio test statistic that is asymptotically distributed as χ^2 with $p_2 - p_1$ degrees of freedom. Along with the AIC and the BIC, the log likelihood was included (along with the degrees of freedom) in the results as well, as a model with a maximum likelihood is considered to fit the data the best. AIC and BIC theories have the same objective: to find the best model via comparison. However, each theory has a different motivation. While the AIC compares models using a measure of similarity of the expected predictive performance, the BIC compares the probabilities that each of the models tested is the true model [44]. The main idea behind the criteria is to compare models based on their maximised log-likelihood values, while penalising them based on the number of parameters. The model with the smallest AIC/BIC value is deemed the best [46].

Additionally, in finding the smallest AIC and BIC values, the model chosen needs to fit the data. For this reason, we again measured R^2, the coefficient of determination [47]. R^2 is a proportionate reduction in the total variation associated with the use of a predictor variable, X. The larger the R^2 value, the more the total variation in Y is reduced by introducing the predictor variable, X. R^2 values occur in the range of $0 \leq R^2 \leq 1$. Values closer to 1 indicate

a better fit. For mixed-effects models, R^2 can be categorised into two types: marginal R^2 and conditional R^2. Marginal R^2 accounts for variance explained by fixed factors:

$$R_M{}^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2} \quad (4)$$

and conditional R^2 is concerned with variance explained by both fixed and random factors [15]:

$$R_C^2 = \frac{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2} \quad (5)$$

where σ_f^2 = the variance calculated from the fixed-effects component; u = the number of random factors in the model; σ_l^2 = the variance component of the lth random factor; and $\left(\sigma_e^2 + \sigma_d^2\right)$ = the sum of the additive dispersion component and the distribution-specific variance.

Once a suitable model was identified and fitted, we then proceeded to check whether the key assumptions of the model were satisfied. These assumptions included (i) linearity, (ii) the homoscedasticity or constancy of the error variance, and (iii) the normality of the errors. Discrepancies between the assumed model and the data could be identified by studying the residuals (also known as the error components). The residuals represent the differences between observed and predicted values per the assumed model. Visual aids, such as residual plots, help identify whether the assumptions of the model have been met. Typically, a good residual plot would be one with an even distribution or symmetry; those that contain distinguishable patterns, such as clustering to one side of the plot, usually indicate a violation of the model's assumptions and warrant a further review of the model (e.g., the transformation of the outcome variable or the exploration of additional variables) [48]. Normal probability plots additionally allowed us to determine the fit of our model.

Using both residual plots and normal probability plots, we could identify any unusual or outlying observations based on large deviations in the observed Y values from the fitted line. Inferences from the model can be potentially influenced by only a few cases in the data. The fitted model may reflect the unusual characteristics of those cases rather than the overall relationship between the dependent and independent variables [49].

2.7. Prediction Accuracy Using Forecasting Errors

Predictions of future events are called forecasts, and the act of making predictions is called forecasting. In attempting to forecast future events, a forecaster must rely on and analyse past data. Past data allow the identification of a pattern which can be extrapolated or extended into the future in order to prepare a forecast. Forecasting techniques rely on the assumption that the patterns which have been identified in the past will continue in the future. Good predictions cannot be expected unless this assumption is valid. That said, all forecasting situations involve a certain degree of uncertainty. There may be an irregular component which may be substantial and cause fluctuations in the data. Hence, we reviewed forecasting errors in an attempt to ascertain whether an irregular component can be so large as to completely invalidate any forecasting technique or whether a forecasting technique might not be capable of accurately predicting the trend, seasonal, or cyclical components of data, thus rendering the technique inappropriate [50].

Several forecast errors can identify whether a forecasting technique does or does not match the pattern of the data. To measure forecasting errors, we used the following notations [51]:

- $\hat{Y}_t$: The forecast generated for the ith patient at time t (i.e., the fitted/predicted value);
- Y_t: The observed value for the ith patient at time t;
- $\overline{Y_t}$: The mean (or centred) value of the outcome measure;
- $e_t = Y_t - \hat{Y}_t$: The forecast error for the ith patient at a particular time t.

The first metric used to determine forecast quality was the mean error (ME), which is simply the average of past errors:

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t \quad (6)$$

This metric captures whether the forecasting process, on average, tends to under-forecast (i.e., ME is positive) or over-forecast (i.e., ME is negative); it is, in fact, a metric of bias. We therefore needed other metrics to assess forecasting accuracy that could capture the proximity between the predictions produced using our model and the actual observed values. An ideal or preferred ME would be as close to zero as possible.

The second metric used was the MAE, which is the absolute value between a forecasted value and an observed value. While the ME can capture the tendency to over- or under-forecast, the MAE tells us how large (or small) a forecasting error we can expect on average. The values for this metric ranges are [0, ∞); the smaller the value, the better. The MAE makes it easier to interpret errors and minimises the weight given to outliers that could skew results. It is expressed as follows:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \quad (7)$$

The third metric used in determining forecasting accuracy was the mean absolute deviation (MAD). MAD uses the absolute error to make sure negative and positive errors add up and conceptualises the amount of error and is robust against outliers. Similar to the preceding metrics, the range of values is [0, ∞), and the smaller the value, the better the result.

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |Y_t - \overline{Y_t}| \quad (8)$$

The fourth metric used to assess forecasting accuracy was the root mean square error (RMSE). This measure squares errors to sum positive and negative ones. The RMSE is used as an estimate of the standard deviation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \quad (9)$$

The values for RMSE occur in the range of [0, ∞). The smaller the RMSE, the better. RMSE measurements are in the same unit as that of the forecast demand. If, for example, we were using kg as the unit, then the RMSE would be measured in kg as well. In our case, our units of measure are mm^2/year (original scale), mm/year (square-root-transformed scale), and % enlargement rate ($\log(Y_i + 1)$ transformation scale).

3. Results

3.1. Feature Extraction and Selection

All 431 features were extracted across all patient data. These extracted features were added to the existing data collected using RegionFinder (i.e., the total area of GA in mm^2/year) and exported into a standard comma-separated values (CSV) file. Using a five-fold cross-validation method, the top 30 features were extracted (Figure 2). A close-up of the feature ranking plot can be found in Figure 3. In the context of feature extraction, this meant that the top 30 most highly ranked features were extracted across all five folds. The objective of the five-fold cross-validation process here was to determine whether the top-ranked features were consistently the same across all five cross-validations, which would confirm the importance of these features relative to the output of GA total area (mm^2/year). For completeness, if there were some features that appeared in one or two cross-validations but not all, those features were still listed as potential candidates for consideration in the modelling process just to ensure that all potential top variables were observed in the model. The final complete list of features tested for modelling can be found in Table 2. The feature "lesion_elongation" was always the top-ranked feature across all five cross-validations, with "lesion_elongation" representing the largest and second-largest principal component axes of lesions. While the clustering variables "early_cluster", "late_cluster", and "lesion_cluster"

did not appear in the top-ranked variables across any of the cross-validations, they were still considered and manually tested given our interest in determining whether there is an association with the new cluster groups and GA progression. The cluster variables were assessed in the traditional univariate linear mixed-effects model style. In a univariate stage, all three clustering variables showed statistical significance with a threshold of $p < 0.05$.

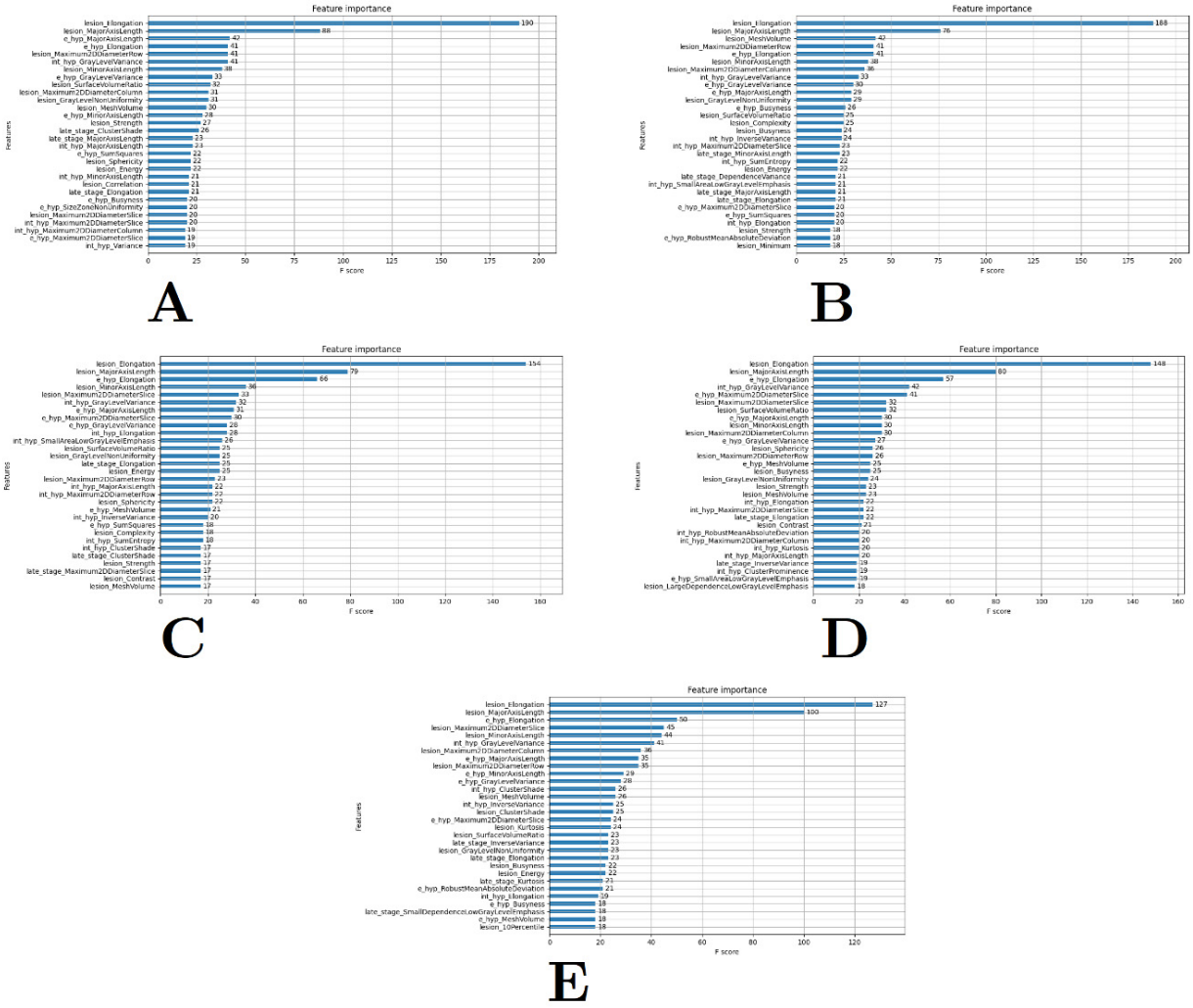


Figure 2. Five-fold cross-validation (**A**–**E**) of top 30 ranked fundus autofluorescence radiomic image features. Most top-ranked features were consistently found across all five cross-validation sets. However, there were slight discrepancies, with some features appearing once or twice in the five-fold tests. For completeness, all variables which appeared across all cross-validations were included in the model assessment process.

These top features were then manually tested in linear mixed-effects models with the original scale (total area mm^2/year) with the random effects of (eye | subject). This originally involved the addition of all the variables of interest identified in Table 2 in a singular model, and the significance of the variables was assessed using the $p < 0.05$ threshold. Then, one by one, variables that were not significant against this threshold were eliminated and the model was rerun. This process continued until all the features tested in the original model were statistically significant. Once these base statistically significant variables had been identified, they were then tested in various combinations, as shown in Table 3. These combinations were observed to see which combined sets of variables produced the most promising modelling results. These combinations were tested against our various modelling and forecasting metrics: R^2_M, R^2_C, Pearson's r, RMSE, ME, MAE, MAD, AIC, BIC, and log likelihood (Table 4). Furthermore, normal Q-Q plots (Figure 4) and

predicted vs. output plots (Figure 5) show that all these various combinations did meet assumptions of normality and had good correlations between expected and observed values. However, the most suitable combination belonged to a model named model 1.2, which contained the variables "lesion_elongation", "lesion_minoraxislength", "lesion_meshvolume", "lesion_contrast", "e_hyp_busyness", "lesion_10percentile", "lesion_sphericity", "lesion_cluster", "e_hyp_sumsquares", "lesion_correlation", "late_stage_clustershade", "lesion_complexity", "late_stage_minoraxislength", "late_stage_dependencevariance", and "late_stage_smalldependencelowgray".

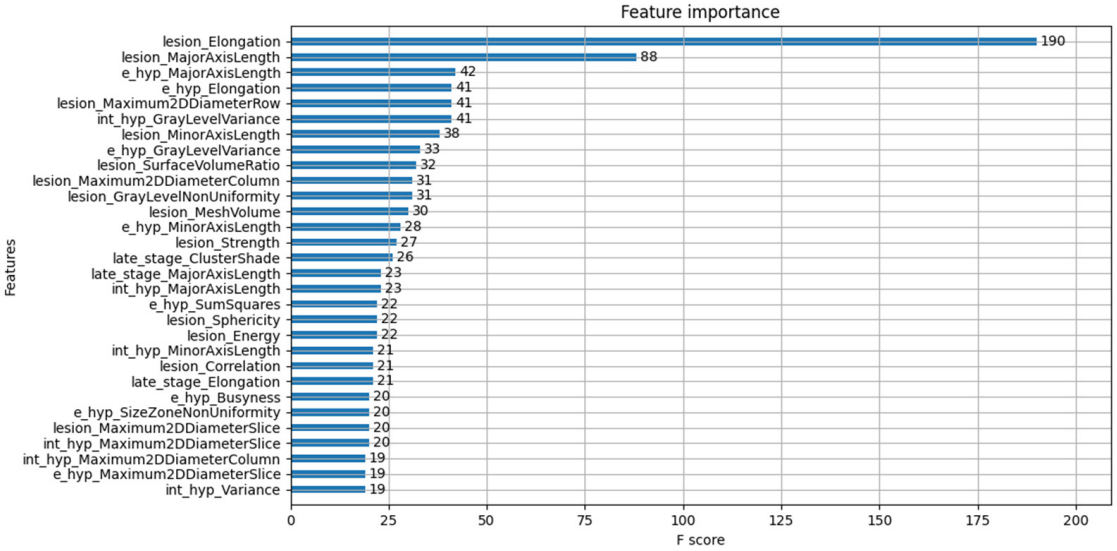


Figure 3. Close-up of the cross-validation 1 feature ranking process. Features beginning with "lesion_" in their name represent lesion-related features, while "e_hyp_" refers to early-stage hyperfluorescence, "int_hyp_" refers to intermediate-stage hyperfluorescence, and "late_hyp_" refers to late-stage hyperfluorescence features. The feature "lesion_elongation" was consistently the top-ranked feature across all five cross-validations.

Table 2. Top features extracted using XGBoost.

Top Features Tested
lesion_elongation
lesion_majoraxislength
e_hyp_majoraxislength
e_hyp_elongation
lesion_maximum2ddiametercolumn
int_hyp_graylevelvariance
lesion_minoraxislength
e_hyp_graylevelvariance
lesion_surfacevolumeratio
lesion_graylevelnonuniformity
lesion_meshvolume
e_hyp_minoraxislength
lesion_strength
late_stage_majoraxislength
int_hyp_majoraxislength
e_hyp_sumsquares
lesion_sphericity
lesion_energy
int_hyp_minoraxislength
lesion_correlation
late_stage_elongation
e_hyp_busyness

Table 2. *Cont.*

Top Features Tested
e_hyp_sizezonenonuniformity
lesion_maximum2ddiameterslice
int_hyp_maximum2ddiameterslice
int_hyp_maximum2ddiametercolumn
e_hyp_maximum2ddiameterslice
int_hyp_variance
late_stage_clustershade
lesion_maximum2ddiameterrow
lesion_complexity
lesion_busyness
int_hyp_inversevariance
late_stage_minoraxislength
int_hyp_sumentropy
late_stage_dependencevariance
int_hyp_smallarealowgraylevelemp
int_hyp_elongation
e_hyp_robustmeanabsolutedeviation
lesion_minimum
e_hyp_meshvolume
int_hyp_clustershade
late_stage_maximum2ddiameterslice
lesion_contrast
int_hyp_robustmeanabsolutedeviat
int_hyp_kurtosis
late_stage_inversevariance
int_hyp_clusterprominence
e_hyp_smallarealowgraylevelempha
lesion_largedependencelowgraylevel
lesion_kurtosis
late_stage_kurtosis
late_stage_smalldependencelowgray
lesion_10percentile

Table 3. Top combinations of features tested for eight linear mixed-effects models for the original scale—GA total area (mm^2/year).

Feature	Model 1.0	Model 1.1	Model 1.2	Model 1.3	Model 1.4	Model 1.5	Model 1.6	Model 1.7
lesion_elongation	x	x	x	x	x	x	x	x
lesion_minoraxislength	x	x	x	x	x	x	x	x
lesion_meshvolume	x	x	x	x	x	x	x	x
lesion_surfacevolumeratio	x	x						
lesion_contrast	x	x	x	x	x	x	x	x
e_hyp_majoraxislength	x	x						
e_hyp_minoraxislength	x	x						
e_hyp_busyness	x	x	x	x	x	x	x	x
lesion_10percentile	x	x	x	x	x	x	x	x
lesion_sphericity	x	x	x	x	x	x	x	x
lesion_cluster		x	x			x	x	
e_hyp_sumsquares			x	x				
lesion_correlation			x	x	x	x		
late_stage_clustershade			x	x	x	x		
lesion_complexity			x	x	x	x		
late_stage_minoraxislength			x	x				
late_stage_dependencevariance			x	x	x	x		
late_stage_smalldependencelowgray			x	x	x	x		
lesion_minimum					x	x		
late_stage_maximum2ddiameterslice					x	x		

Table 4. Results from the eight linear mixed-effects models from Table 3 ranked in order of most suitable to least suitable.

Model	R^2_M	R^2_C	*r*	*p*-Value	RMSE	ME	MAE	MAD	AIC	BIC	logLik	DF
Model 1.2	0.83	0.96	0.981	<0.001	1.32	-7.3×10^{-15}	0.94	0.999	2084.93	2169.97	−1022.46	20
Model 1.1	0.83	0.95	0.979	<0.001	1.35	-1.4×10^{-14}	0.97	1.047	2124.56	2192.77	−1046.28	16
Model 1.5	0.83	0.95	0.981	<0.001	1.32	-1.2×10^{-16}	0.94	0.995	2094.51	2179.55	−1027.25	20
Model 1.6	0.82	0.95	0.978	<0.001	1.40	-4.3×10^{-15}	0.99	1.063	2131.48	2186.90	−1052.74	13
Model 1.3	0.75	0.94	0.979	<0.001	1.38	-6.8×10^{-15}	0.97	0.969	2215.01	2292.22	−1089.50	18
Model 1.4	0.76	0.94	0.979	<0.001	1.37	1.9×10^{-16}	0.97	0.992	2215.14	2292.35	−1089.57	18
Model 1.0	0.76	0.94	0.978	<0.001	1.42	-8.6×10^{-15}	1.01	1.027	2257.33	2317.57	−1114.67	14
Model 1.7	0.74	0.94	0.976	<0.001	1.46	-5.1×10^{-15}	1.02	1.002	2260.57	2307.90	−1119.28	11

R^2_M: Marginal R^2; R^2_C: Conditional R^2; *r*: Pearson's correlation coefficient with the accompanying *p*-value; RMSE: Root mean square error; ME: Mean error; MAE: Mean absolute error; MAD: Mean absolute deviation; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; logLik: Log likelihood; DF: Degrees of freedom.

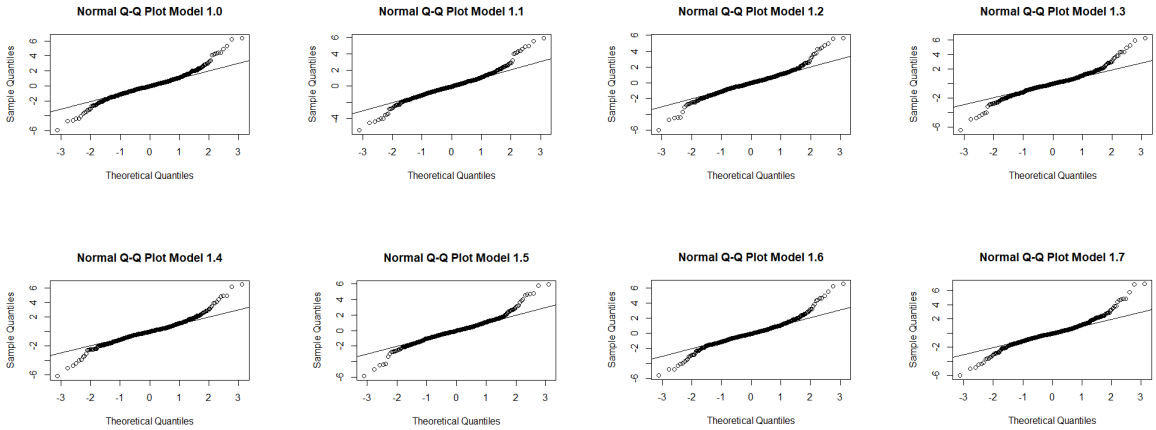


Figure 4. Normal Q-Q plots for linear mixed-effects models used to test various combinations of features of interest. All plots have very similar outputs and are generally well established across the line except for the tail ends. The assumption of normality, irrespective of the combination of variables, was supported by these results.

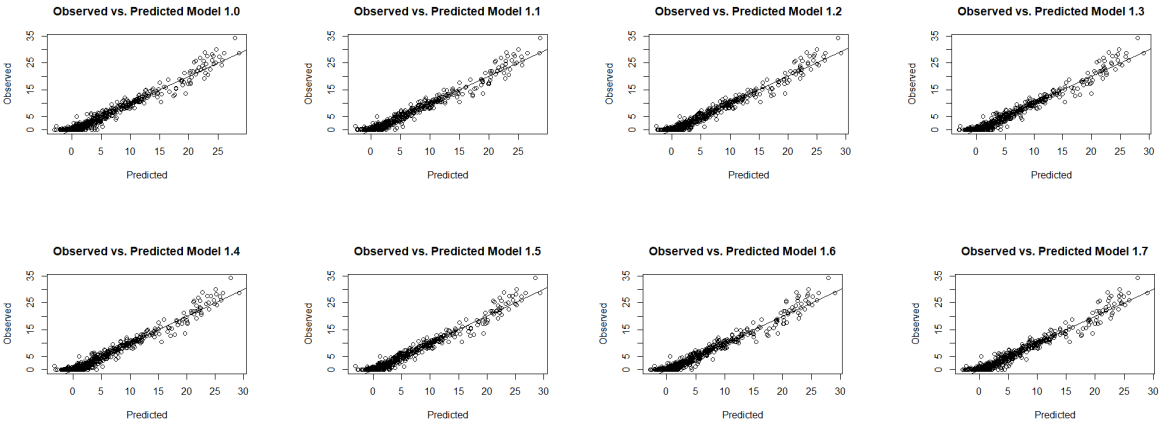


Figure 5. Observed vs. predicted plots for linear mixed-effects models used to test various combinations of features of interest. All plots have very similar outputs and show high correlations between the outputs and predictions.

This combination of variables had one of the highest conditional and marginal R^2 values (R^2_M = 0.83, R^2_C = 0.96), a high correlation coefficient (r = 0.981; $p < 0.001$]), some of the smallest forecasting errors (RMSE = 1.32, ME = -7.3×10^{-15}, MAE = 0.94, and MAD = 0.999), and the lowest criterion values (AIC = 2084.93, BIC = 2169.97, and log likelihood = −1022.46). Therefore, in the succeeding sections, these variables were further tested

using (1) the original, square-root-transformed, and $\log(Y_i + 1)$-transformed outcomes, and (2) all of these model outcomes were tested using another five-fold cross-validation method similar to the one applied to the linear mixed-effects model. The transformations were tested to see if they improved the fit of the variables to the model.

3.2. Model Selection

The models described above were used to test which of the top featured variables would be most suitable when put into a linear mixed-effects model. These top features were determined using the original scale (total area mm^2/year) as the outcome of interest with the random effects of (eye | subject). Initially, assessment of the most suitable variables was performed using the entirety of the dataset. However, to minimise any overfitting, the features identified in Section 3.1 were then evaluated using the original scale using the cross-validation approach. The process was then repeated for the two additional outcomes of square-root transformation and $\log(Y_i + 1)$ transformation. The features identified in model 1.2 were tested against the new and transformed outcomes to see if they maintained any statistical significance with the transformation. Most features still maintained their statistical significance in the square-root-transformed model (i.e., the model named 1.2sq), with only "late_stage_clustershade" and "lesion_complexity" losing statistical significance with the square-root-transformed outcome. However, with the $\log(Y_i + 1)$ transformation (i.e., the model named 1.2log), the variable list reduced somewhat, with only "lesion_minoraxislength", "lesion_meshvolume", "lesion_contrast", "e_hyp_busyness", "lesion_10percentile", "lesion_sphericity", "lesion_cluster", and "late_stage_smalldependencelowgray" maintaining their statistical significance with the threshold of $p < 0.05$. A complete list of the final variables used across the three different linear mixed-effects models can be found in Table 5. The models 1.2sq and 1.2log also underwent five-fold cross-validation for a complete comparison (Table 6).

Table 5. The different features used in the original scale (model 1.2), along with its square-root- (model 1.2sq) and $\log(Y_i + 1)$ (model 1.2log)-transformed forms.

Feature	Model 1.2	Model 1.2sq	Model 1.2log
lesion_elongation	x	x	
lesion_minoraxislength	x	x	x
lesion_meshvolume	x	x	x
lesion_contrast	x	x	x
e_hyp_busyness	x	x	x
lesion_10percentile	x	x	x
lesion_sphericity	x	x	x
lesion_cluster	x	x	x
e_hyp_sumsquares	x	x	
lesion_correlation	x	x	
late_stage_clustershade	x		
lesion_complexity	x		
late_stage_minoraxislength	x	x	
late_stage_dependencevariance	x	x	
late_stage_smalldependencelowgray	x	x	x

At first glance, the performance of the square-root-transformation model quantitatively appears to supersede those of the original scaled and log-transformed models. Additionally, the model produces the most tight-fitting normal Q-Q plots relative to the other two models (Figure 7) and has a good residual plot (Figure 13). However, upon closer inspection of Table 6, it will be noticed that, save for the forecasting metrics of RMSE, ME, MAE, and MAD, all the other metrics have *identical* results across all five folds of cross-validations. This suggests that the square-root-transformed model is most likely an overfitted model [13]. In our earlier publication, we demonstrated the effect of the square-root transformation on GA growth areas and progression modelling in the absence of features. The transformation had a flattening effect and a tendency to "tighten" the data points. The overfitting may

be the result of this effect, and thus a square-root-transformed model most likely is not suitable. Furthermore, we can see with the correlation plot of predicted and observed outputs (Figure 10) that there is a curvature at the lower tail end of the plot, suggesting that a perfectly linear relationship between predicted vs. output values is absent. Interestingly, while the marginal and conditional R^2, the correlation coefficient (r), and the AIC/BIC/log-likelihood metrics were consistently the same across all five folds, the varying metrics of RMSE, ME, MAE, and MAD were considerably larger relative to the same values of the original scaled model, 1.2. Thus, given the identical results produced across the five-fold cross-validation, along with the predicted vs. output plot not behaving in the expected manner, the square-root transformation was deemed inappropriate in this study.

Table 6. Results for the original scale (model 1.2), the square-root-transformed scale (model 1.2sq), and $\log(Y_i + 1)$ (model 1.2log) across five cross-validations.

		R^2_M	R^2_C	r	p-Value	RMSE	ME	MAE	MAD	AIC	BIC	logLik	DF
Cross-Validation 1	**Original Area**	0.83	0.95	0.98	<0.001	1.35	0.10	0.93	2.58	1729.33	1809.85	−844.67	20
	Square-Root Area	0.89	0.96	0.93	<0.001	6.15	3.37	3.54	2.92	330.47	407.01	−147.24	18
	Log(Y_i + 1)	0.89	0.96	0.88	<0.001	5.87	3.78	3.86	2.62	62.86	115.10	−18.43	13
Cross-Validation 2	**Original Area**	0.85	0.95	0.96	<0.001	1.39	−0.07	1.05	3.86	1717.88	1798.30	−838.94	20
	Square-Root Area	0.89	0.96	0.95	<0.001	6.49	3.60	3.73	2.52	330.47	407.01	−147.24	18
	Log(Y_i + 1)	0.89	0.95	0.92	<0.001	7.74	4.84	4.89	3.11	126.84	179.37	−50.42	13
Cross-Validation 3	**Original Area**	0.83	0.95	0.98	<0.001	1.52	0.25	1.09	2.85	1712.43	1793.19	−836.22	20
	Square-Root Area	0.89	0.96	0.94	<0.001	6.33	3.89	4.02	3.08	330.47	407.01	−147.24	18
	Log(Y_i + 1)	0.88	0.95	0.91	<0.001	8.66	5.49	5.54	3.14	109.04	161.41	−41.52	13
Cross-Validation 4	**Original Area**	0.84	0.96	0.97	<0.001	1.44	−0.07	1.08	4.05	1701.23	1781.55	−830.61	20
	Square-Root Area	0.89	0.96	0.94	<0.001	6.94	3.93	4.11	2.78	330.47	407.01	−147.24	18
	Log(Y_i + 1)	0.89	0.95	0.90	<0.001	7.23	4.11	4.21	2.24	124.57	176.81	−49.28	13
Cross-Validation 5	**Original Area**	0.85	0.96	0.98	<0.001	1.50	0.26	1.05	4.25	1731.18	1812.08	−845.59	20
	Square-Root Area	0.89	0.96	0.93	<0.001	6.84	3.82	3.95	2.93	330.47	407.01	−147.24	18
	Log(Y_i + 1)	0.89	0.95	0.88	<0.001	8.04	4.66	4.73	2.68	118.31	170.48	−46.15	13

R^2_M: Marginal R^2; R^2_C: Conditional R^2; r: Pearson's correlation coefficient with the accompanying p-value; RMSE: Root mean square error; ME: Mean error; MAE: Mean absolute error; MAD: Mean absolute deviation; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; logLik: Log likelihood; DF: Degrees of freedom.

For the log-transformed variables, the marginal and conditional R^2 values were as promising as the square-root transformation. Additionally, the values did differ between cross-validations, suggesting consistency in outcomes but no overfitting. The AIC/BIC and log likelihood values were also very promising, as they presented the smallest values out of all the models, which usually would suggest that the log-transformed model would be the most promising. The Q-Q plot (Figure 8) and the residual plot (Figure 14) also suggested that the residuals were normally distributed. However, the forecasting errors for the log transformation were the largest out of the three modelling types. Furthermore, the correlation plot for the log-transformed model (Figure 11) produced very similar patterns between predicted and observed values to the square-root-transformed model.

As a result, the original scaled model, 1.2, was deemed to be the most appropriate (represented in Equation (10), where $\beta_1, \beta_2 \ldots, \beta_{15}$ are the fixed coefficients for variables $X_1, X_2, \ldots, X_{15}$ highlighted in model 1.2 and $Z_1\gamma_1 + Z_2\gamma_2$ represent the random effects):

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{15} X_{15} + Z_1\gamma_1 + Z_2\gamma_2 + \epsilon \quad (10)$$

Generally, the quantitative values were very good, with high values for the conditional and marginal R^2 as well as the correlation coefficient. The forecasting errors were smallest with the original scaled model as compared to the transformed models. The average R^2_M for the original scaled model was 0.84 ± 0.01. For R^2_C, the average was 0.95 ± 0.002, the r average was 0.97 ± 0.006, the RMSE was 1.44 ± 0.06, the ME was 0.09 ± 0.15, the

MAE was 1.04 ± 0.06, the MAD was 3.52 ± 0.67, the AIC was 1718.41 ± 11.08, the BIC was 1798.99 ± 11.20, and the log likelihood was −839.21 ± 5.54. While the AIC/BIC and log likelihood were highest for the original scaled model, this was overshadowed by the consistency of the visualisation results. The normal Q-Q plot (Figure 6) may not have been as tight fitting as for the square-root- and log-transformed models, but it was still in line with the assumptions of normality, and, coupled with the residual plots (Figure 12), this further validated that the assumptions of normality were met. Furthermore, the original scaled model was the only one that produced visually correct correlation plots (Figure 9), with the predicted and observed values having a strong linear relationship.

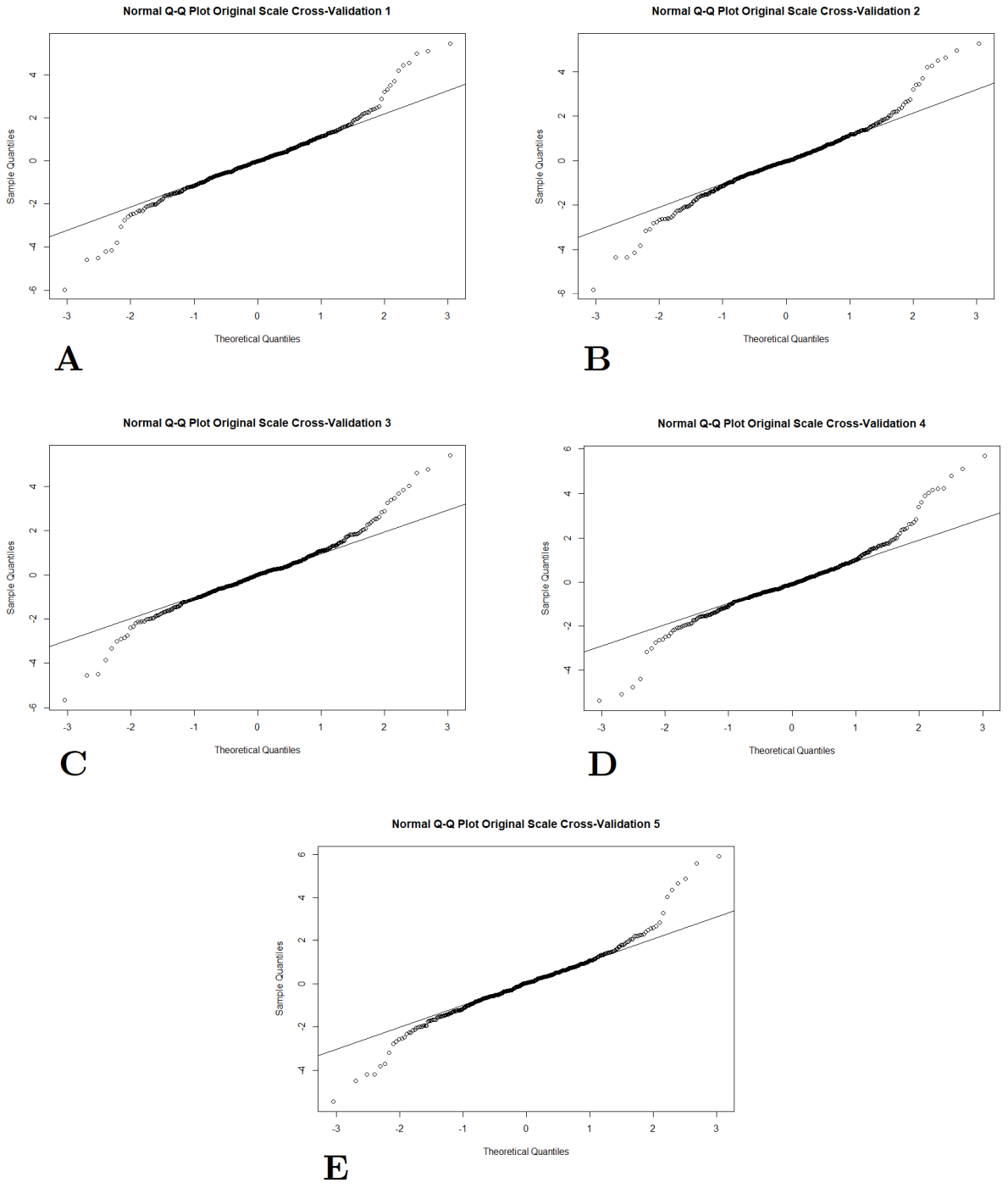


Figure 6. Normal Q-Q plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the original scale of mm^2/year. The plots show a general adherence to the assumption of normality.

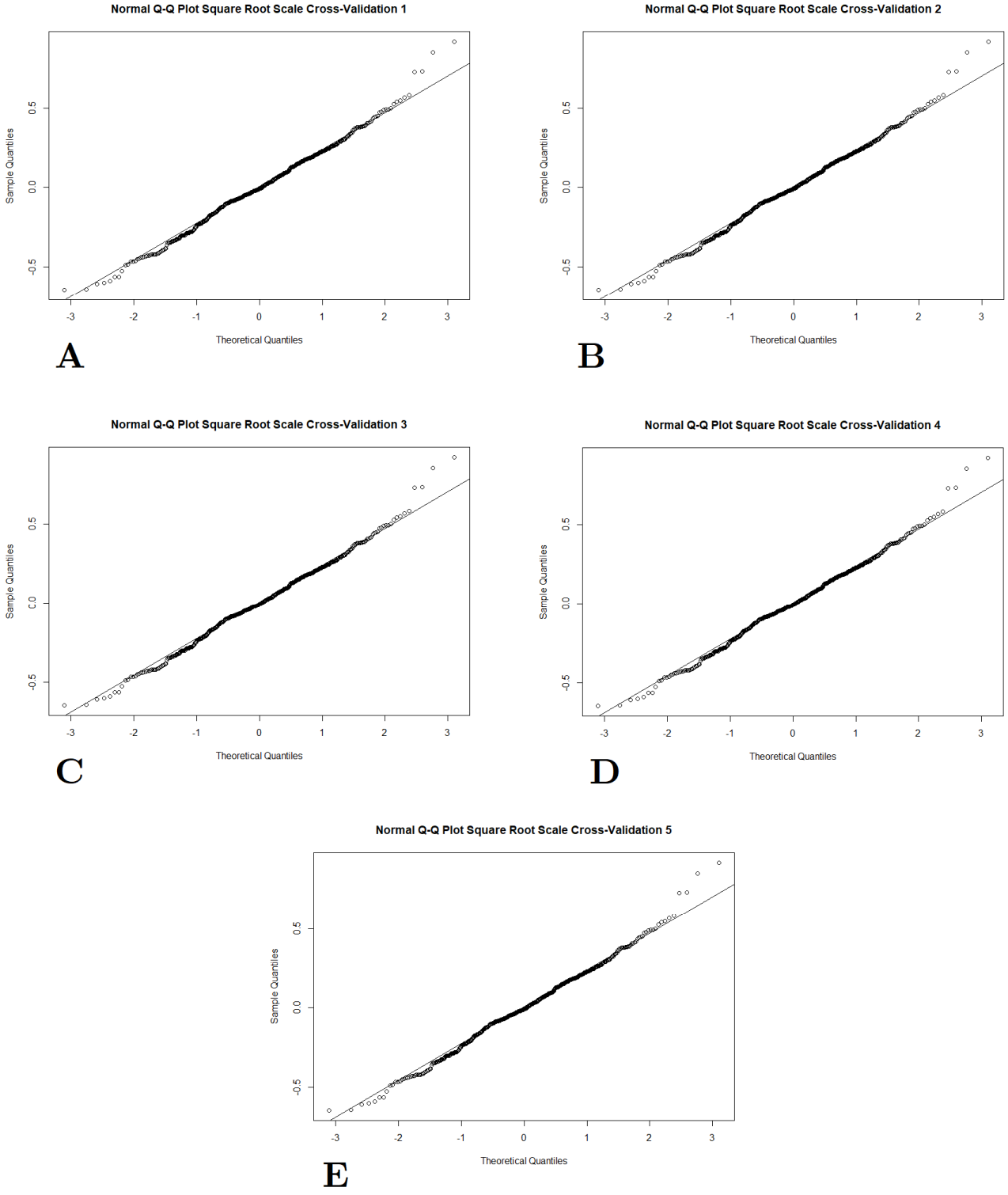


Figure 7. Normal Q-Q plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the square-root scale of mm/year. The plot shows a much tighter fit of residuals around the 45-degree line.

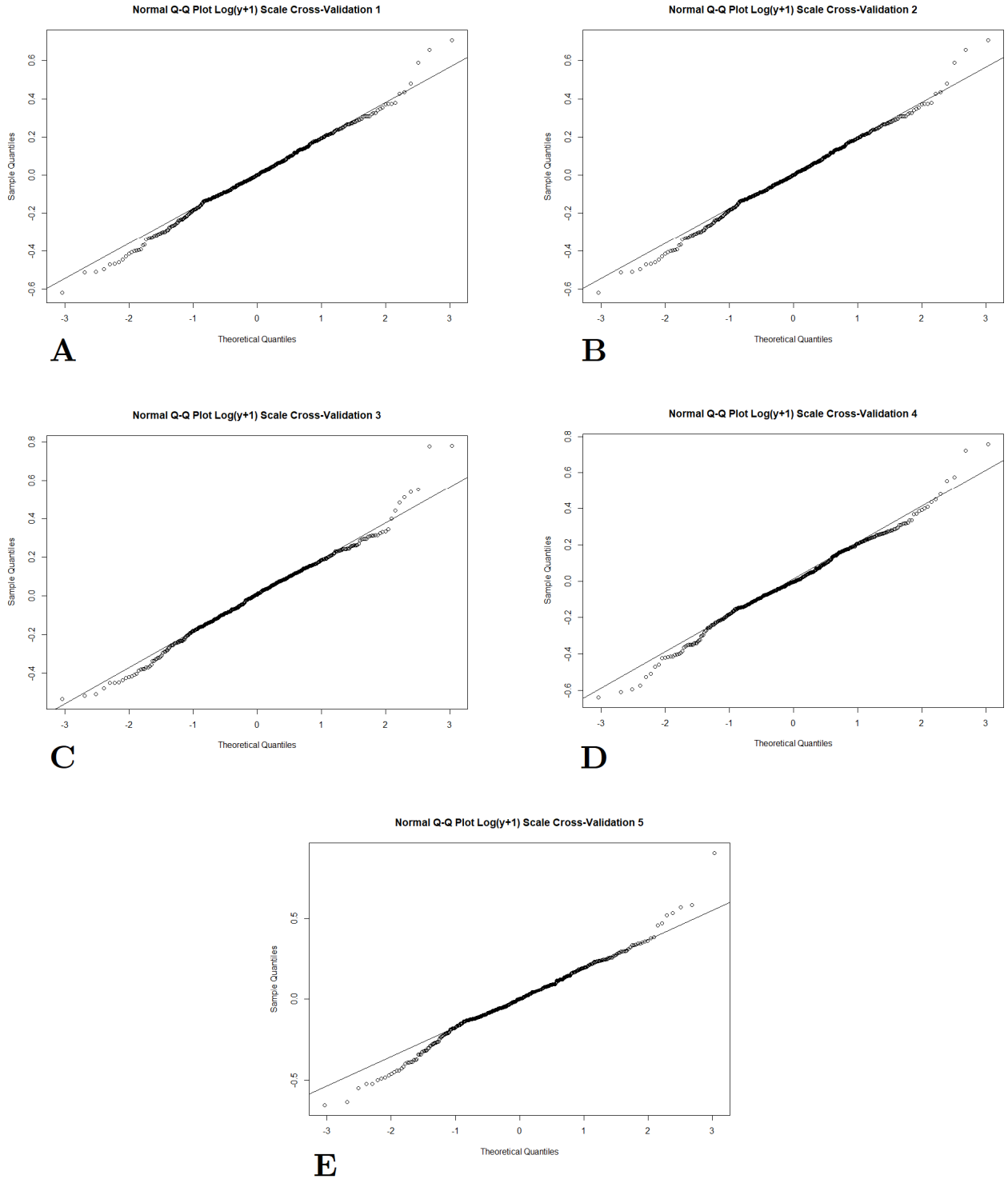


Figure 8. Normal Q-Q plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the $\log(Y_i + 1)$ scale of enlargement rate/year. Generally, there was a good adherence to assumptions of normality.

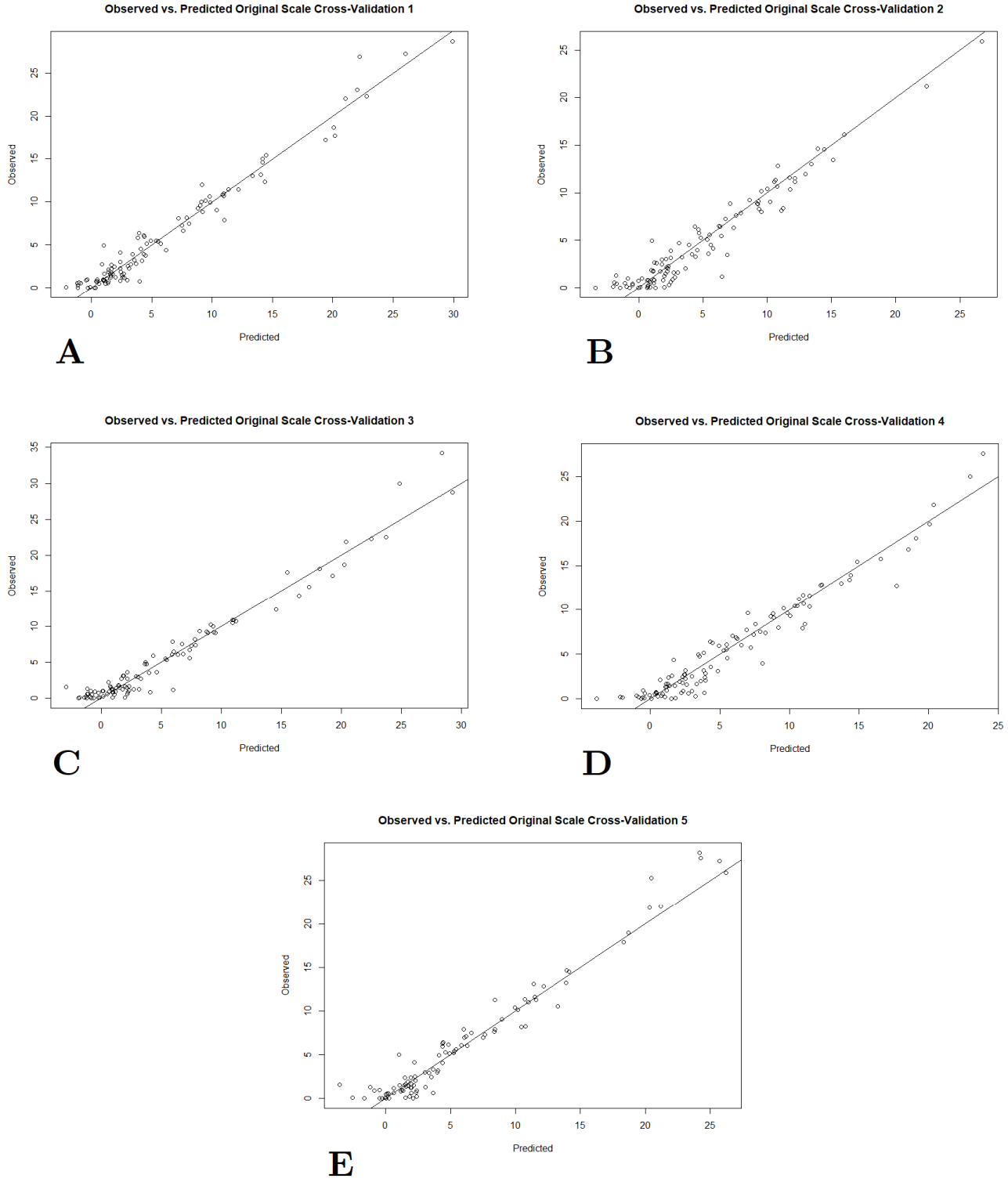


Figure 9. Predicted vs. observed plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the original scale of mm^2/year. There were minimal differences between the predicted and observed values, with a strong linear relationship.

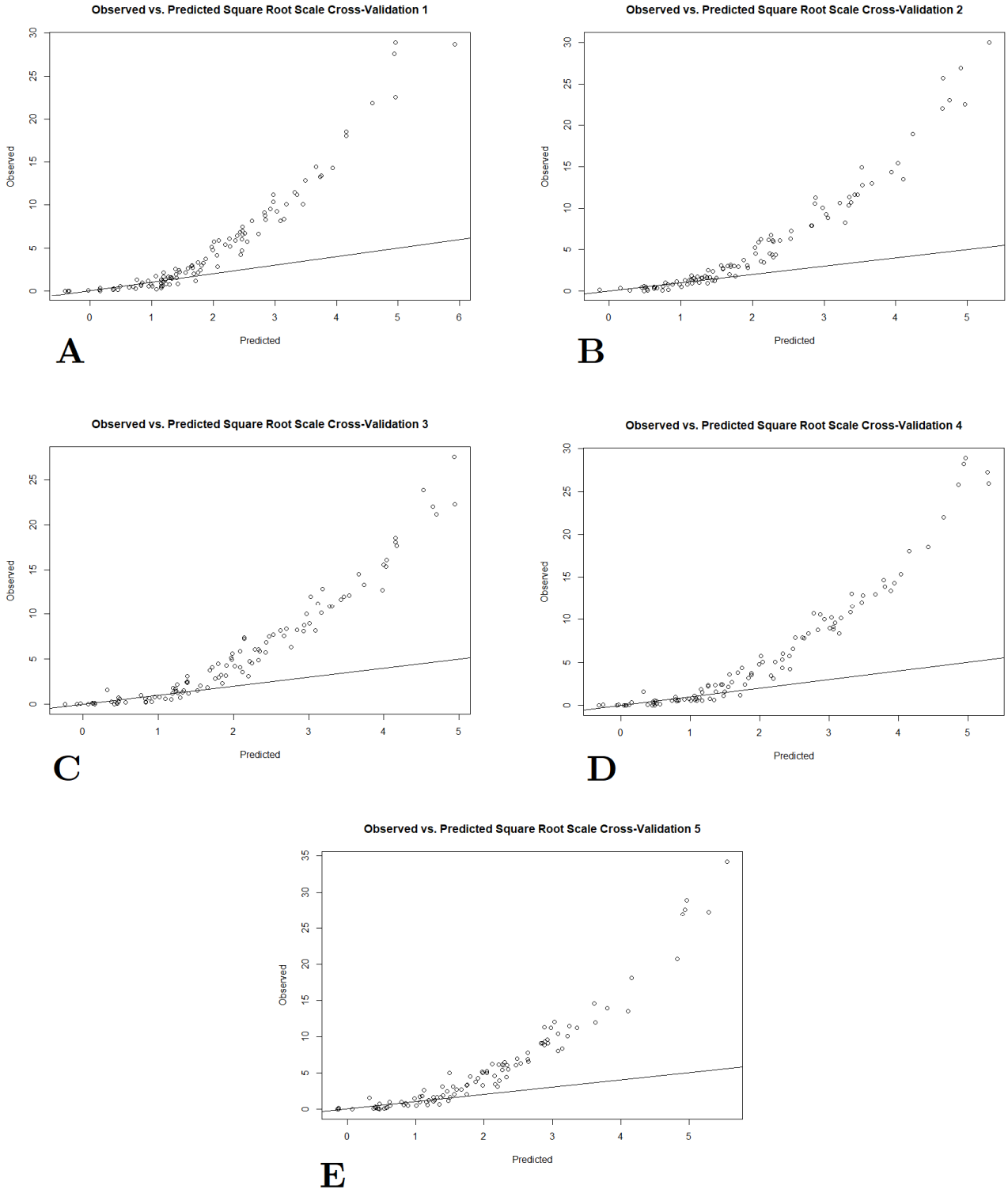


Figure 10. Predicted vs. observed plots for five-fold (**A**–**E**) cross-validation of the linear mixed-effects model with the square-root scale of mm/year. There is a curvature at the lower tail end, and the relationship is not as strong as that of the original scaled model.

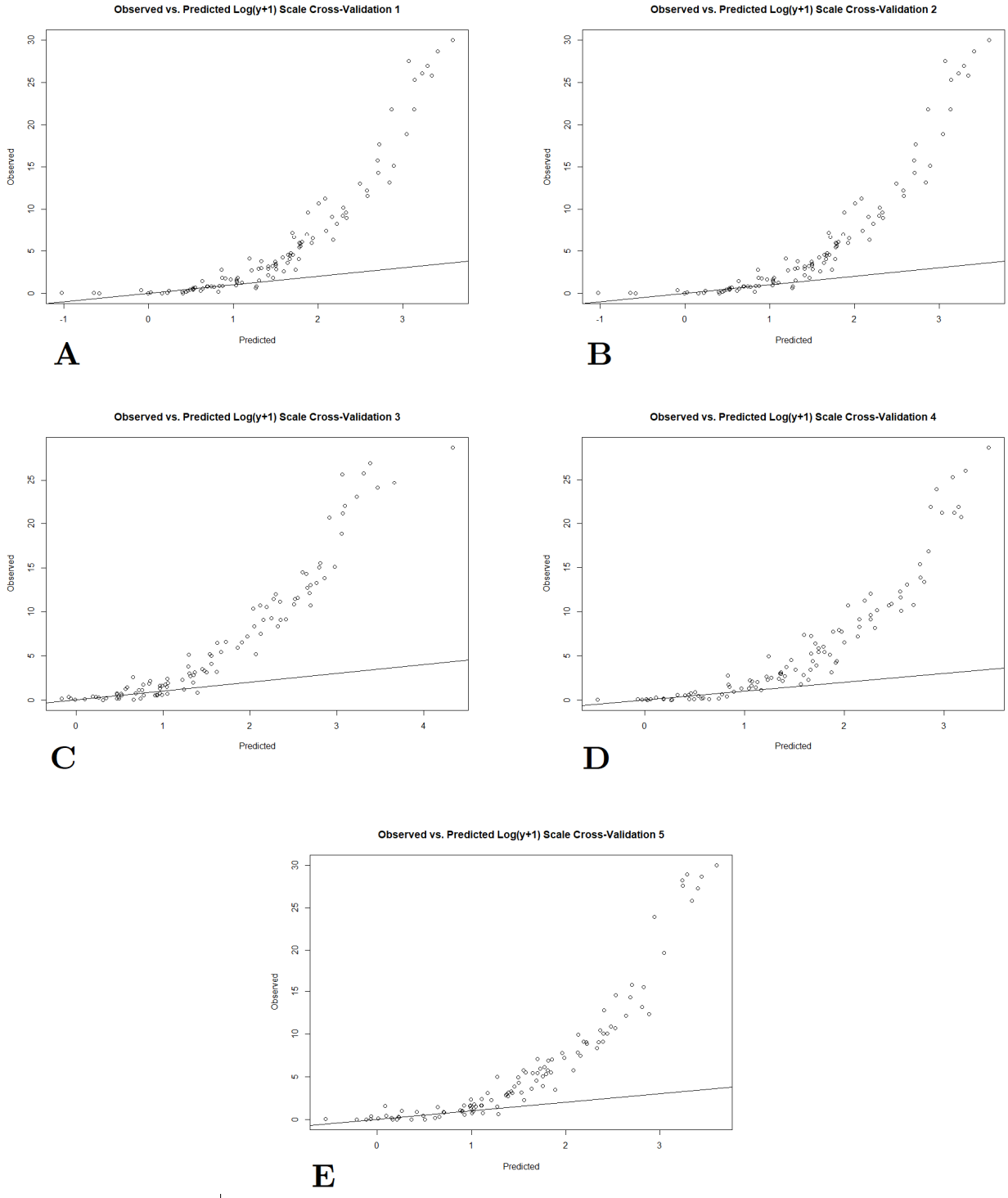


Figure 11. Predicted vs. observed plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the $\log(Y_i + 1)$ scale of enlargement rate/year. There is a curvature at the lower tail end, and the relationship is not as strong as that of the original scaled model.

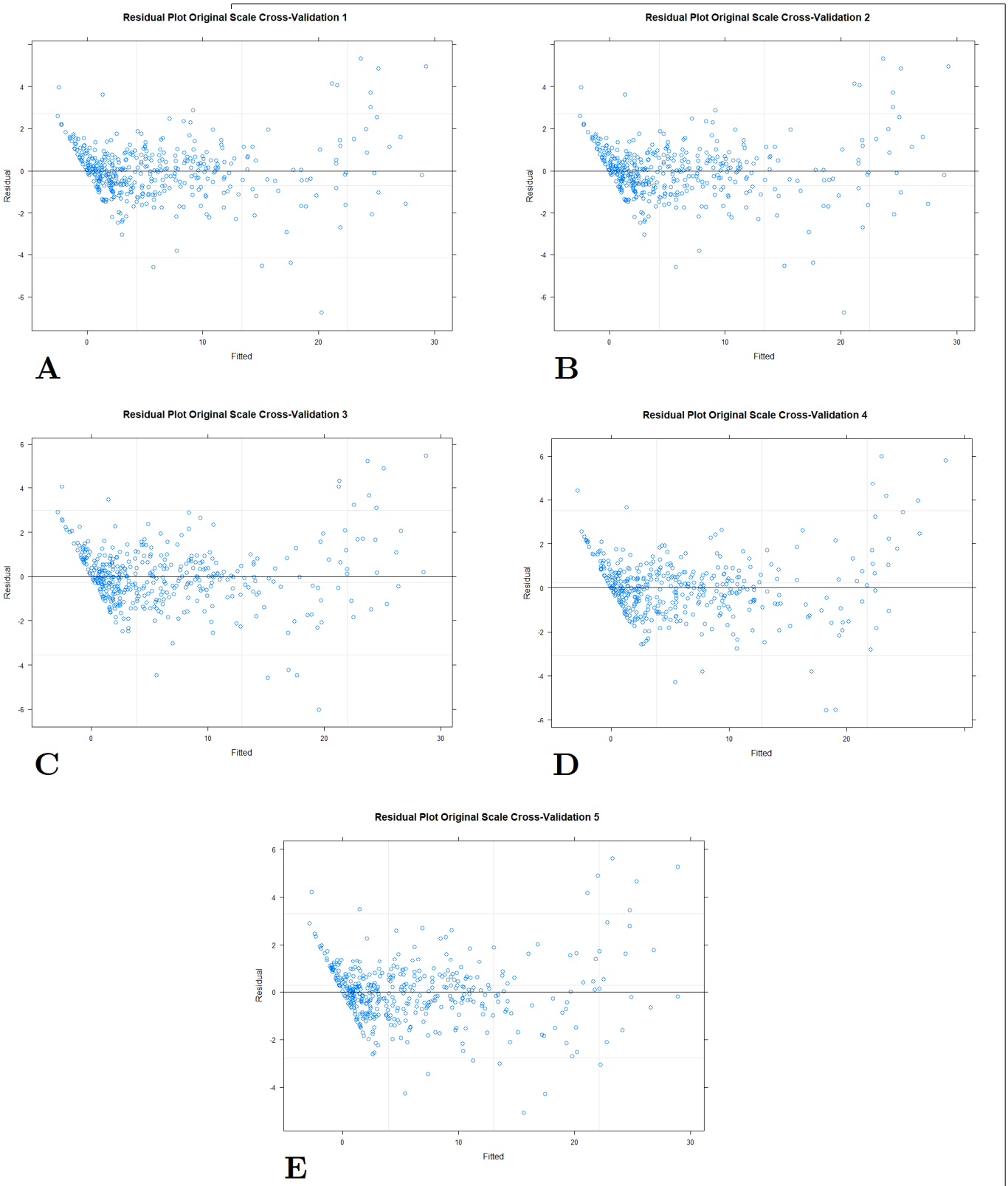


Figure 12. Residual plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the original scale of mm^2/year. Generally, there is an even distribution of residuals, although a slight cluster appears on the left-hand side.

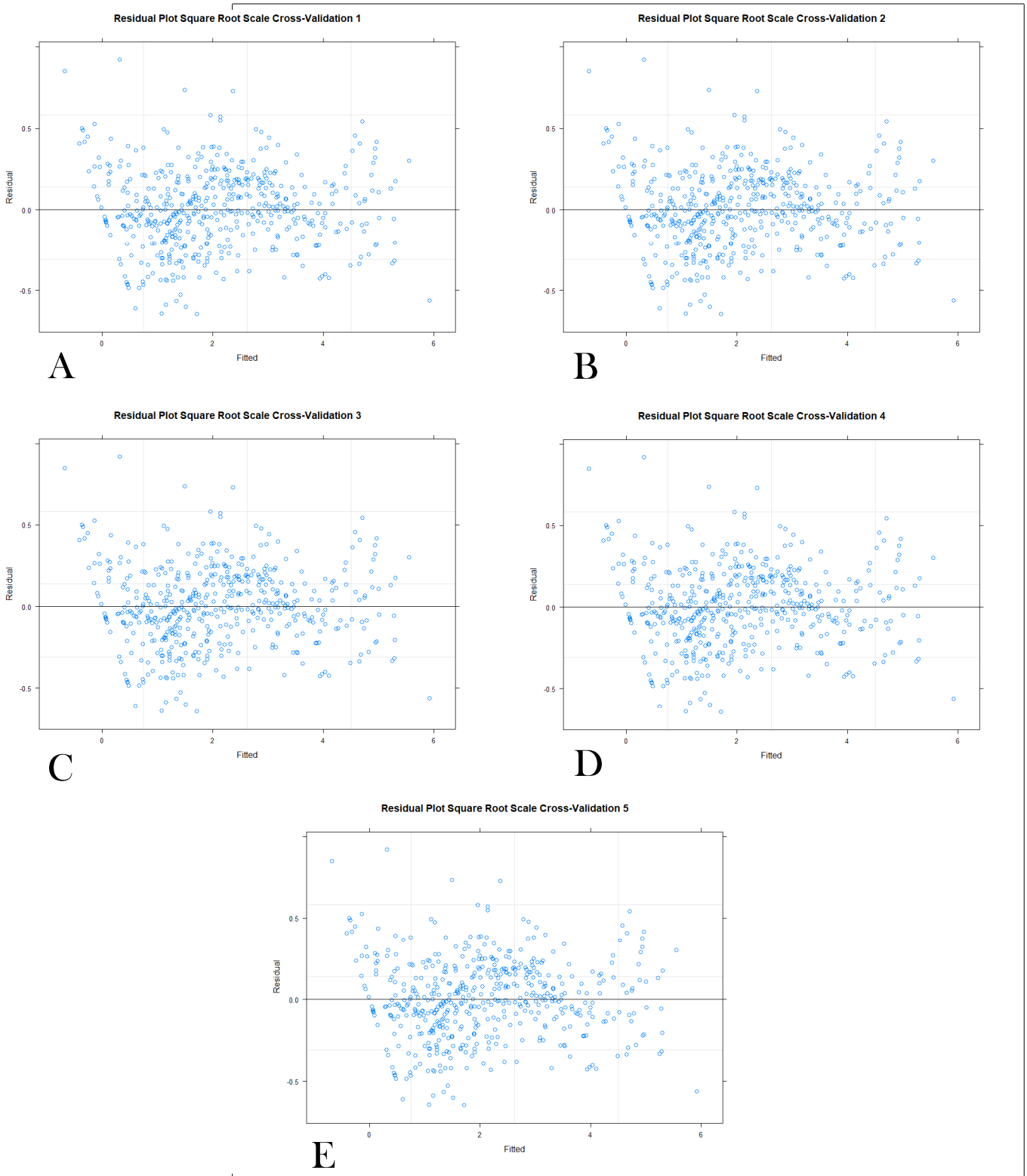


Figure 13. Residual plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the square-root scale of mm/year. Generally, there is an even distribution of residuals, and the slight cluster on the left-hand side is spread out.

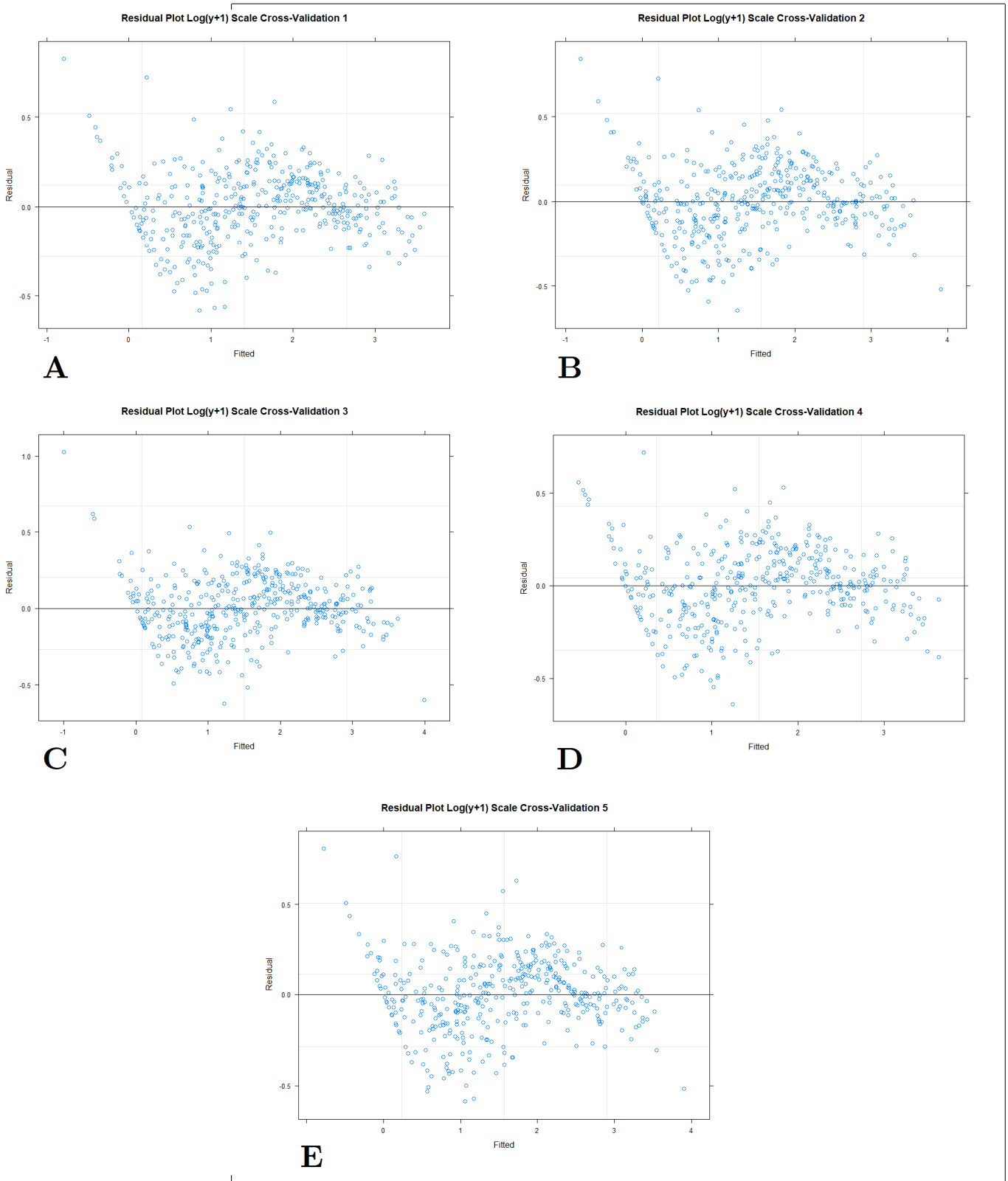


Figure 14. Residuals plots for five-fold cross-validation (**A**–**E**) of the linear mixed-effects model with the $\log(Y_i + 1)$ scale of enlargement rate/year. Generally, there is an even distribution of residuals, and the slight cluster on the left-hand side is spread out.

4. Discussion

4.1. Overall Findings

This paper describes the evaluation of the linear mixed-effects model using FAF-based image features as a prediction model for GA growth. The choice of the linear mixed-effects model stemmed from several factors. Firstly, in our previous studies, it was confirmed that the linear model, representing the high growth part of the progression of GA, offers the best description of trends relative to other regression models, such as the quadratic model without the linear term and the logarithmic model. Furthermore, the mixed-effects model is a hierarchical model that is suitable for time-series data, such as measurements used for GA progression analysis.

The mixed-effects model can also account for changes in slopes (i.e., variability in patient progression) by incorporating random effects. In the models presented, *subject* and *eye* were considered to be the most appropriate random effects to account for changes relative to each individual while also accounting for inter-eye correlation. The sigmoidal function was hypothesised and investigated in [13] with the limited data available and remains a subject for future research. In this study, the linear mixed-effects model was used to describe the peak growth portion (the middle part) of the sigmoidal function. This is because current available clinical datasets are best described by the linear part of the growth curve. However, future extension of this research will require new data (i.e., nascent GA estimates to explain the lower tail end and additional follow-up to explain the upper tail end). The sigmoidal hypothesis was tested using several case studies with the most data points. However, with an expanded real-time dataset, future explorations could include the sigmoidal mixed-effects model.

The features used for the linear mixed-effects model were image features extracted from FAF images using radiomics and the binarised masks created. The binarised masks included lesions, as well as the various stages of hyperfluorescence development, including early-stage, intermediate-stage, and late-stage hyperfluorescence [27]. A total of 107 radiomic features were extracted per feature, which meant a total of 428 radiomic feature extractions. These were coupled with cluster-based categories determined in our prior work on cluster analysis [27]. The variable "lesion_cluster" had three categories: preliminary lesion formation, intermediary lesion formation, and complete lesion formation. The variable "early_cluster" also had three categories: early-stage hyperfluorescence complete coverage, early-stage hyperfluorescence partial coverage, and early-stage hyperfluorescence proximal coverage. The variable "late_cluster" had its own three categories, including late-stage hyperfluorescence droplet scatter, late-stage hyperfluorescence halo scatter, and late-stage hyperfluorescence halo. Intermediate-stage clustering groups were not included, as the cluster groups were not distinguishable. With these three additional variables, the total assessable features were 431. First, XGBoost was utilised to identify the top highly ranked features. These features were then evaluated in the linear mixed-effects model (with the outcome as total area mm^2/year) in a variety of combinations. The combination of variables that boded well against the multitude of metrics was "lesion_elongation", "lesion_minoraxislength", "lesion_meshvolume", "lesion_contrast", "e_hyp_busyness", "lesion_10percentile", "lesion_sphericity", "lesion_cluster", "e_hyp_sumsquares", "lesion_correlation", "late_stage_clustershade", "lesion_complexity", "late_stage_minoraxislength", "late_stage_dependencevariance", and "late_stage_smalldependencelowgray".

Here, a quick explanation of the aforementioned variables is presented. Lesion elongation refers to the largest and second-largest principal component axes of lesions. Lesion minor axis length is the second-largest length of axis for lesions. Lesion mesh volume refers to the total volume of lesions calculated based on a triangle mesh. Lesion contrast refers to the local intensity variation of lesions. Early-stage hyperfluorescence busyness measures the degree of change between early hyperfluorescence pixels and neighbouring pixels. Lesion 10th percentile refers to the 10th percentile of voxels in lesions. Lesion sphericity refers to the measure of roundness of lesions. Lesion cluster is the late-stage clustering category [27]. Early-stage hyperfluorescence sum of squares refers to the intensity of neighbouring pixels

around early hyperfluorescence. Lesion correlation measures the coarseness of lesion textures. Lesion cluster shade measures uniformity or skewness in gray-level occurrences in lesions. Lesion complexity determines if there are any drastic changes in gray-level intensities within lesions. Late-stage hyperfluorescence minor axis length is the second-largest length of axis for late hyperfluorescence. Late-stage hyperfluorescence-dependence variance is the measure of variances within late hyperfluorescence regions. Finally, the late-stage hyperfluorescence small dependence low gray-level emphasis variable measures lower gray-level values [26].

The majority of features found suitable for GA progression modelling were features around lesion-based parameters. Perhaps this is not unusual given that the model presented focuses on the linear portion of the sigmoidal function, rather than the whole sigmoidal function itself. It would be assumed that, as the disease is at its peak progression stage, more lesion-based parameters would be more intuitive in predicting the direction and progression of the disease.

4.2. Testing Models with Different Transformations

The variables were tested in a five-fold cross-validation approach in three different types of linear mixed-effects models: (1) the original scaled total GA area of mm^2/year, (2) the square-root-transformed area of mm/year, and (3) the $\log(Y_i + 1)$-transformed area of enlargement rate/year. Using a combination of quantitative and visualisation techniques, the original scaled linear mixed-effects model proved to be the most suitable for the current data.

In this work, we investigated the popular square-root transformation (1) when assessing the most suitable growth rate in the absence of features and (2) when assessing the growth rate with image features. In [13], it was confirmed that the square-root transformation did not have an impact on the assumptions of normality, as the linear model with and without the square-root transformation had residuals that were normally distributed. In this study, the normal Q-Q plot for the square-root transformation at first appeared to be a tighter and better fit than the original scale. The residual plots also showed a relatively uniform distribution of residuals, which is indicative of a good fit. However, visual assessments aside, the square-root-transformation model's quantitative metrics were questionable. Despite running the model using the five-fold cross-validation process, most of the model diagnostic metrics remained identical, including the marginal and conditional R^2, the Pearson's correlation coefficient r, the AIC/BIC, and the log likelihood. This would suggest that the square-root transformation might actually create an overfitted model and is thus most likely to be rigid and inflexible in its application outside of the current available dataset. Furthermore, the square-root transformation produced poorer outcomes in terms of forecasting errors, having much greater values for RMSE, ME, MAE, and MAD as compared to the original scale. The analysis for the square-root transformation is a good illustration of why a combination of techniques needs to be utilised when assessing any model; a greater range of assessment methods improves confidence in the choice of model most suitable for the available data.

This study also investigated the $\log(Y_i + 1)$ transformation. It appears that this alternative has not been explored previously in the literature. The use of the log transformation with an additive constant is more suitable in GA progression analysis. This is due to the small sizes of some lesions which, if subjected to the traditional log transformation, would be converted to negative values. However, the inclusion of the constant "1" prevents this, ensuring that all transformed values remain positive.

The $\log(Y_i + 1)$ transformation at first appeared to be a promising alternative to the more popular square-root transformation. As with the square-root model, it produced normal Q-Q plots that had a tight fit around the 45-degree line which validated that the residuals were normally distributed. It had an evenly distributed residual plot, further validating that the assumptions of normality for the model had been met. Advantageously, the $\log(Y_i + 1)$ transformation did not have identical model fit values across its five-fold

cross-validations to the square-root transformation. There were slight differences between each iteration, but these were consistent enough for confidence in the results. Furthermore, amongst all the models, the $\log(Y_i + 1)$ model had the smallest AIC, BIC, and log-likelihood values. If these values were used in isolation of other metrics of interest, the $\log(Y_i + 1)$-transformed model would have been considered the top model in this study. However, the $\log(Y_i + 1)$ transformation had two primary downfalls: (1) it had an identical predicted vs. observed plot to that of the square-root transformation, which did not support confidence in its predictive abilities, and (2) it had the largest forecasting errors of the three modelling types. Thus, while the $\log(Y_i + 1)$ transformation may be promising and could be explored with models that include additional features beyond FAF image features, in the context of this study, the $\log(Y_i + 1)$ transformation was not appropriate.

If the $\log(Y_i + 1)$ transformation were to produce improved results in future studies, there would be the question of how to interpret its meaning. In this study, the $\log(Y_i + 1)$ transformation outcome was referenced as the enlargement rate per year. However, this may still create confusion in a clinical setting. A way to gain clarity in future studies would be to (1) create a model with the $\log(Y_i + 1)$ transformation and then (2), once the prediction had been made by the model, the prediction could be reverse-transformed using the formula $\exp(Y_i) - 1$. The final output would still be the familiar total area of mm^2/year, and the transformation could be used safely to develop a highly predictive model. If, in the current study, the $\log(Y_i + 1)$ model had been the most capable of the three modelling types, this approach of reverse transformation would have been used to predict GA as total area mm^2/year.

4.3. The Most Suitable Model

The original scaled linear mixed-effects model was the most suitable model in this study. It was the model that ticked all the boxes, including having good quantitative model fit metrics and small forecasting errors and meeting the assumptions of normality. Although the residual plots for the original scaled model were near perfect, some minor clustering/clumping to the left-hand side of the plots was notable. This might be suggestive of further modifications being required to improve the performance of the model. Transformations are typically one approach to improve the spread of residuals; however, this study confirmed that transformations did not yield the benefits that would be normally expected. Other possible considerations for model improvement could be the augmentation of the current available data. For example, in this case, we were limited to the use of predominantly FAF image features. But GA is a complex disease, and image features alone do not cover every facet of this progressive ocular condition. Thus, the current model and its residual plots could be improved with the addition of new image or clinical variables. Alternatively, there may be interaction terms that have not been explored in this case that could also help improve the model. It should be noted that interaction terms can also introduce complexities and are not always readily interpretable. Essentially, a small clumping of residuals suggests that something is missing and needs to be included in the model for completeness, and, given the current singular-modality nature of the available dataset, some necessary improvements are to be expected.

The cluster variables "early_stage", "late_stage", and "lesion_stage" were explored in the analyses, even though they did not appear in the top highly ranked features when using XGBoost. Their inclusion was still assessed to see if these newly defined cluster variables can be used to predict GA progression. Given that the variables did not appear in the top-ranked features, they were manually and individually assessed in a traditional univariate mixed-effects model fashion to see if, in isolation of other features, there was any statistical significance with any of the cluster variables.

The early-stage and late-stage hyperfluorescence and lesion cluster variables all showed statistical significance in a univariate mixed-effects model analysis. However, when placed in a model with other covariates, only lesion clustering maintained its significance. This could possibly suggest that the hyperfluorescence clusters can, in part,

explain GA manifestation and progression. However, these clusters may be more suitable for explaining earlier stages of GA. Here, the lesion cluster variable was included in the model, suggesting that lesion clustering is the most useful out of the clustering variables to explain GA progression *once GA lesions have appeared*. It is hypothesised that early-stage and late-stage hyperfluorescence clustering variables could instead be used to model *when a GA lesion will appear*, and thus these clusters may be more useful for the lower tail end of the sigmoidal curve or nascent GA stages. Thus, the clustering variables could be reconsidered when we assess the sigmoidal mixed-effects model in future extensions of the work reported in this study.

4.4. Comparisons with the Literature

The closest comparable model is one based on the paper by Pfau et al. (2019) [5]. In their paper, they assessed shape descriptors in 296 eyes from 201 patients using FAF and IR images and a linear mixed-effects model. These shape descriptors included lesion area, perimeter, and circularity, all of which were normalised using the square-root transformation. Pfau and colleagues used the coefficient of determination, R^2, as their outcome measure, and also used the Pearson's r correlation coefficient. They assessed the model in two clinically relevant scenarios: (1) where the patient was previously unknown and (2) where prediction was made using previous observations from the patient. They also conducted their analyses by stratifying lesion sizes into quartiles. In their first scenario model, they found that the square-root of the lesion area, the square root of the perimeter, the focality index, the hyperfluorescence phenotype, the square root of the circularity, and the minimum and maximum ferret diameters created a model with an R^2 value of 0.244. For their Scenario 2 model, which included the same variables but with prior knowledge of the patient, the R^2 value increased to 0.391.

Künzel et al. (2020) focused on the association of VRQoL and visual function/structural biomarkers in GA in 87 patients with FAF, IR, and SD-OCT imaging [10]. Similar to Pfau et al. (2019), they too square-root-transformed their GA size. They fitted a linear regression model to the complete dataset at baseline using LASSO regression and compared it to the linear mixed-effects model, in which the patient was the random effect. They produced an R^2 of 0.32.

In comparison, the linear mixed-effects model produced the following results across the five-fold cross-validations. The average R^2_M value for the original scaled model was 0.84 ± 0.01. For R^2_C, the average was 0.95 ± 0.002, the r average was 0.97 ± 0.006, the RMSE was 1.44 ± 0.06, the ME was 0.09 ± 0.15, the MAE was 1.04 ± 0.06, the MAD was 3.52 ± 0.67, the AIC was 1718.41 ± 11.08, the BIC was 1798.99 ± 11.20, and the log likelihood was -839.21 ± 5.54. The metrics used in this study were more comprehensive in scope as compared to previously published papers. The transformations were also more extensive than in other studies, and the focus regarding transformations was not solely on the square-root transformation. The log transformation was considered in depth, and the transformation was conducted considering small GA areas, which was not an objective in previous studies surveyed by the authors. Both the marginal and conditional R^2 values in this study are higher than those presented in the literature for similar linear mixed-effects models. Furthermore, the model presented here, in its original and untransformed scale, proved to produce smaller forecasting errors than the transformed scale. Therefore, while future recommendations are made in the following paragraph to improve the current dynamics and coverage of the linear mixed-effects model presented here, it has still proven to be a much more accurate model when compared with others in the literature.

4.5. Limitations and Future Work

In this paper, the linear mixed-effects model was tested extensively using radiomics for feature extraction. The information elicited from the data could also be used to eventually develop a sound and consistent biophysical model. In [12], we briefly touched on biophysical models and the advantage of using such models in medicine. This study demonstrated

the value of using biophysical models as well as statistical models. We showed how purely statistical models can introduce uncertainty in performance, with variations in the covariates and parameters used. We have also seen how overfitting can easily occur, even with models that initially seem to tick all the boxes for model diagnostics. These variabilities and uncertainties are due to the reliance of statistical and ML models on sample data for training rather than on physical properties of the disease itself and their relationships. The data-dependent parameters for training a statistical prediction model will change when the training dataset changes. We assume that the training sample is representative of the test dataset. The cross-validation process is designed with the aim of minimising overfitting and lowering uncertainty.

Problems relating to training and testing are generally not issues for biophysical models (apart from model calibration, in some cases, if required), but biophysical models also have some limitations. A disadvantage associated with biophysical models is that their development can take a considerable amount of time based on anthropic observations and hypothesis testing, whereas more timely predictions are possible with standard statistical models. A correctly formulated biophysical model is based on fundamental theories and assumptions and can apply to all datasets, whether large or small, without the need for training data every time. Most biophysical models have only a small number of covariates, as the derivation of non-linear models is not a trivial exercise. Statistical models, in contrast, may have a very large number of covariates and parameters that are beyond the scope of many biophysical models. While the derivation of biophysical models was not in the scope of this thesis, future work could involve the development of biophysical GA progression models to complement or enhance machine learning models.

One of the first objectives for future work is to address the extraction of new image features. The features in this study were extracted using XGBoost against the original scaled total area (in units of $mm^2/year$). While these features were modified accordingly based on model diagnostics when used for the square-root- and $\log(Y_i + 1)$-transformed equivalent models, the features extracted were not specifically run with these outcomes in mind, and thus other features that were not included in the top-ranked features might have been more relevant to these outcomes. However, this limitation is another demonstration of how future models could be augmented by a greater contribution from biophysical modelling, which also helps with the "explainability" of model outputs. The authors have found that the use of convolution filtering in machine learning has, in the past, enabled more physically plausible feature extraction based on local shape or texture in an image, and this would complement time-based mixed-effects models based on first- and second-order statistics [28,52].

There are also other complexities with statistical datasets that are often ignored. An interesting complexity is referred to as the suppression effect. Using the definition by Conger (1974) [53], a suppressor variable is defined as one which increases the validity of another variable simply by its inclusion in a multivariate model. This effect may be due to dataset complexities, such as a small and unbalanced sample size, or the presence of interactions which have not yet been explored. This was briefly shown in the analyses here when, in isolation, the cluster variables showed statistical significance but lost their significance or changed their significance in the presence of other variables.

A strength of this study was the identification of potentially invariant image features associated with GA progression that allowed decoupling of the prediction model from dependence on new statistical samples for training data. For example, the use of an image feature, such as lesion elongation, as a metric for GA progression, is an attribute associated with a physical model rather than a statistical model. The mixed-effects model also introduced time dependence, which is not often associated with machine learning analysis. In some future situations, it may be valuable to incorporate clinical data explicitly into the model for more comprehensive analysis. Clinical data could be used in combination with image-based features to improve our understanding of GA progression. Because potential confounders may not be readily identified by imaging alone, the inclusion of

clinical variables may reveal valuable additional information (relating to age, gender, or comorbidities).

5. Conclusions

Radiomic features were extracted from FAF images which described the shape, size, and intensities of GA features, including lesions, early-stage hyperfluorescence, intermediate-stage hyperfluorescence, and finally late-stage hyperfluorescence. These features were then ranked, and the top 30 most important features were selected for the modelling process using XGBoost, a machine learning algorithm. There were a number of reasons for choosing XGBoost over more traditional statistical techniques for ranking variables. Firstly, the parallel computing process of XGBoost means that many analyses are performed simultaneously, which contributes to an increase in computational speed and efficiency. Secondly, XGBoost has the ability to take weak learners and convert them into strong learners and thus produce more accurate results. Thirdly, as XGBoost is a gradient-descent decision tree, it sifts through and "splits" the data and it will automatically remove redundant variables (i.e., those which are highly correlated), thus preventing the need to check for multicollinearity at later stages of analysis. These features are also reasons why XGBoost has become a popular ML algorithm.

Following the ranking of the image features, the features were then tested in three different linear mixed-effects models: (1) the original scaled outcome model, (2) the square-root-transformed outcome model, and (3) the log-transformed [i.e., $\log(\boldsymbol{Y_i} + 1)$] outcome model. While, in [13], we showed that the popular square-root transformation had similar residual outcomes to that of the original scaled linear model, the transformation was tested here again to determine if the transformation improved the model fitting in the presence of image features. The square-root transformation at first appeared to have been a better fit when combined with the image features. However, with large forecasting errors, discrepancies in predicted vs. observed values, and model fit values that did not change at all with each iteration of the cross-validation, it was concluded that the square-root transformation was not suitable and most likely was producing an overfitted model.

The log transformation appeared to be a positive alternative at first. It was stipulated that the lack of success or minimal use of this transformation in the past could have been because the $\log(\boldsymbol{Y_i})$ transformation was being used instead of the log transformation with the additive constant of "1". The former log transformation can create negative values with small GA areas. However, with the inclusion of 1, this problem is avoided. Unlike its square-root-transformed counterpart, values were slightly different between each iteration of the cross-validation, and model fit values, such as AIC, BIC, and log likelihood, were promising. However, the log transformation also suffered from forecasting errors and had larger variability in its predictions as compared to the original scaled linear mixed-effects model. Nonetheless, the new $\log(\boldsymbol{Y_i} + 1)$ transformation should be considered in future studies and compared to the more commonly used square-root transformation. The main disadvantage of this new log transformation is its interpretability. Previous publications have used the enlargement rate per year for the standard log transformation, and this publication followed suit. However, with the new log transformation, to make results more interpretable, one could run a reverse transformation following the model prediction. This would mean that one would first conduct and run the model with the $\log(\boldsymbol{Y_i} + 1)$ transformation. Once the prediction has been produced, then one could apply a reverse transform using $\exp(\boldsymbol{Y_i}) - 1$.

Author Contributions: Conceptualisation, J.A. and K.B.; methodology, J.A.; software, J.A.; validation, J.A. and K.B.; formal analysis, J.A.; investigation, J.A. and K.B.; writing—original draft preparation, J.A.; writing—review and editing, J.A. and K.B.; visualisation, J.A.; supervision, K.B.; project administration, K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the Royal Victorian Eye and Ear Hospital (HREC: project no. 95/283H/15; approval date: 28 March 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Holz, F.G.; Bindewald-Wittich, A.; Fleckenstein, M.; Dreyhaupt, J.; Scholl, H.P.; Schmitz-Valckenberg, S.; Group, F.A.-S. Progression of geographic atrophy and impact of fundus autofluorescence patterns in age-related macular degeneration. *Am. J. Ophthalmol.* **2007**, *143*, 463–472. [CrossRef]
2. Jeong, Y.J.; Hong, I.H.; Chung, J.K.; Kim, K.L.; Kim, H.K.; Park, S.P. Predictors for the progression of geographic atrophy in patients with age-related macular degeneration: Fundus autofluorescence study with modified fundus camera. *Eye* **2014**, *28*, 209–218. [CrossRef] [PubMed]
3. Klein, R.; Meuer, S.M.; Knudtson, M.D.; Klein, B.E. The epidemiology of progression of pure geographic atrophy: The Beaver Dam Eye Study. *Am. J. Ophthalmol.* **2008**, *146*, 692–699. [CrossRef]
4. Lindblad, A.S.; Lloyd, P.C.; Clemons, T.E.; Gensler, G.R.; Ferris, F.L., 3rd; Klein, M.L.; Armstrong, J.R.; Age-Related Eye Disease Study Research Group. Change in area of geographic atrophy in the Age-Related Eye Disease Study: AREDS report number 26. *Arch. Ophthalmol.* **2009**, *127*, 1168–1174. [CrossRef]
5. Pfau, M.; Lindner, M.; Goerdt, L.; Thiele, S.; Nadal, J.; Schmid, M.; Schmitz-Valckenberg, S.; Sadda, S.R.; Holz, F.G.; Fleckenstein, M. Prognostic Value of Shape-Descriptive Factors for the Progression of Geographic Atrophy Secondary to Age-Related Macular Degeneration. *Retina* **2019**, *39*, 1527–1540. [CrossRef]
6. Liefers, B.; Colijn, J.M.; González-Gonzalo, C.; Verzijden, T.; Wang, J.J.; Joachim, N.; Mitchell, P.; Hoyng, C.B.; van Ginneken, B.; Klaver, C.C.W.; et al. A Deep Learning Model for Segmentation of Geographic Atrophy to Study Its Long-Term Natural History. *Ophthalmology* **2020**, *127*, 1086–1096. [CrossRef]
7. Niu, S.; de Sisternes, L.; Chen, Q.; Rubin, D.L.; Leng, T. Fully Automated Prediction of Geographic Atrophy Growth Using Quantitative Spectral-Domain Optical Coherence Tomography Biomarkers. *Ophthalmology* **2016**, *123*, 1737–1750. [CrossRef]
8. Pfau, M.; Möller, P.T.; Künzel, S.H.; von der Emde, L.; Lindner, M.; Thiele, S.; Dysli, C.; Nadal, J.; Schmid, M.; Schmitz-Valckenberg, S.; et al. Type 1 Choroidal Neovascularization Is Associated with Reduced Localized Progression of Atrophy in Age-Related Macular Degeneration. *Ophthalmol. Retin.* **2020**, *4*, 238–248. [CrossRef]
9. Schmidt-Erfurth, U.; Bogunovic, H.; Grechenig, C.; Bui, P.; Fabianska, M.; Waldstein, S.; Reiter, G.S. Role of deep learning quantified hyperreflective foci for the prediction of geographic atrophy progression. *Am. J. Ophthalmol.* **2020**, *216*, 257–270. [CrossRef]
10. Künzel, S.H.; Möller, P.T.; Lindner, M.; Goerdt, L.; Nadal, J.; Schmid, M.; Schmitz-Valckenberg, S.; Holz, F.G.; Fleckenstein, M.; Pfau, M. Determinants of Quality of Life in Geographic Atrophy Secondary to Age-Related Macular Degeneration. *Investig. Ophthalmol. Vis. Sci.* **2020**, *61*, 63. [CrossRef] [PubMed]
11. Pfau, M.; Emde, L.v.d.; Dysli, C.; Möller, P.T.; Thiele, S.; Lindner, M.; Schmid, M.; Rubin, D.L.; Fleckenstein, M.; Holz, F.G.; et al. Determinants of cone- and rod-function in geographic atrophy: AI-based structure-function correlation. *Am. J. Ophthalmol.* **2020**, *217*, 162–173. [CrossRef]
12. Arslan, J.; Benke, K.K. Progression of Geographic Atrophy: Epistemic Uncertainties Affecting Mathematical Models and Machine Learning. *Transl. Vis. Sci. Technol.* **2021**, *10*, 3. [CrossRef] [PubMed]
13. Arslan, J.; Benke, K.K.; Samarasinghe, G.; Sowmya, A.; Guymer, R.H.; Baird, P.N. Model Structure Uncertainty in the Characterization and Growth of Geographic Atrophy. *Transl. Vis. Sci. Technol.* **2021**, *10*, 2. [CrossRef] [PubMed]
14. Ying, G.-s.; Maguire, M.G.; Glynn, R.; Rosner, B. Tutorial on Biostatistics: Linear Regression Analysis of Continuous Correlated Eye Data. *Ophthalmic Epidemiol.* **2017**, *24*, 130–140. [CrossRef] [PubMed]
15. Nakagawa, S.; Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **2013**, *4*, 133–142. [CrossRef]
16. Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C.M.; Davidson, R.I.; Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **2011**, *8*, 833–835. [CrossRef] [PubMed]
17. Korte, A.; Vilhjalmsson, B.J.; Segura, V.; Platt, A.; Long, Q.; Nordborg, M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **2012**, *44*, 1066–1071. [CrossRef] [PubMed]
18. Gondro, C.; van der Werf, J.; Hayes, B. *Genome-Wide Association Studies and Genomic Prediction*; Springer: New York, NY, USA; Heidelberg, Germany; Dordrecht, The Netherlands; London, UK, 2013.
19. Demidenko, E. *Mixed Models: Theory and Application with R*; John Wiley & Sons: Hoboken, NJ, USA, 2013.

20. Zhou, X.; Carbonetto, P.; Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* **2013**, *9*, e1003264. [CrossRef] [PubMed]
21. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [CrossRef] [PubMed]
22. Li, M.; Liu, X.; Bradbury, P.; Yu, J.; Zhang, Y.-M.; Todhunter, R.J.; Buckler, E.S.; Zhang, Z. Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **2014**, *12*, 73. [CrossRef]
23. Rizzo, S.; Botta, F.; Raimondi, S.; Origgi, D.; Fanciullo, C.; Morganti, A.G.; Bellomi, M. Radiomics: The facts and the challenges of image analysis. *Eur. Radiol. Exp.* **2018**, *2*, 36. [CrossRef]
24. Burham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.
25. Burnham, K.P. *Multimodel Inference: Understanding AIC Relative Variable Importance Values*; Colorado State University: Fort Collins, CO, USA, 2015.
26. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
27. Arslan, J.; Benke, K. Automation of Cluster Extraction in Fundus Autofluorescence Images of Geographic Atrophy. *Appl. Biosci.* **2023**, *2*, 384–405. [CrossRef]
28. Arslan, J.; Samarasinghe, G.; Sowmya, A.; Benke, K.K.; Hodgson, L.A.B.; Guymer, R.H.; Baird, P.N. Deep Learning Applied to Automated Segmentation of Geographic Atrophy in Fundus Autofluorescence Images. *Transl. Vision Sci. Technol.* **2021**, *10*, 2. [CrossRef]
29. Quinto, B. *Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More*; Elahi, I., Ed.; Apress: New York, NY, USA, 2020.
30. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
31. Pinheiro, J.; Bates, D. *Mixed-Effects Models in S and S-PLUS*; Springer: New York, NY, USA, 2000.
32. Dingemanse, N.J.; Dochtermann, N.A. Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *J. Anim. Ecol.* **2013**, *82*, 39–54. [CrossRef]
33. Edwards, L.J.; Muller, K.E.; Wolfinger, R.D.; Qaqish, B.F.; Schabenberger, O. An R^2 Statistic for Fixed Effects in the Linear Mixed Model. *Stat. Med.* **2008**, *27*, 6137–6157. [CrossRef]
34. Zhang, Z.; Ersoz, E.; Lai, C.-Q.; Todhunter, R.J.; Tiwari, H.K.; Gore, M.A.; Bradbury, P.J.; Yu, J.; Arnett, D.K.; Ordovas, J.M.; et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **2010**, *42*, 355–360. [CrossRef]
35. Cheng, J.; Edwards, L.J.; Maldonado-Molina, M.M.; Komro, K.A.; Muller, K.E. Real Longitudinal Data Analysis for Real People: Building a Good Enough Mixed Model. *Stat. Med.* **2010**, *29*, 504–520. [CrossRef]
36. Brown, H.; Prescott, R. *Applied Mixed Models in Medicine*; John Wiley & Sons Ltd.: West Sussex, UK, 1999.
37. Ho, R.K.W.; Hu, I. Flexible modelling of random effects in linear mixed models—A Bayesian approach. *Comput. Stat. Data Anal.* **2008**, *52*, 1347–1361. [CrossRef]
38. Schutzenmeister, A.; Piepho, H.-P. Residual analysis of linear mixed models using a simulation approach. *Comput. Stat. Data Anal.* **2012**, *56*, 1405–1416. [CrossRef]
39. Proust, C.; Jacqmin-Gadda, H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput. Methods Programs Biomed.* **2005**, *78*, 165–173. [CrossRef]
40. Baayen, R.H.; Davidson, D.J.; Bates, D.M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **2008**, *59*, 390–412. [CrossRef]
41. Chen, Z.; Dunson, D.B. Random Effects Selection in Linear Mixed Models. *Biometrics* **2003**, *59*, 762–769. [CrossRef]
42. Bates, D.; Maechler, M. *Package 'lme4': Linear Mixed Model Implementation*; R Core Team: Vienna, Austria, 2010.
43. Brader, H.S.; Pistilli, M.; Ying, G.-S.; Maguire, M.G. Early Progression of Geographic Atrophy in the Complications of Age-related macular degeneration Prevention Trial (CAPT). *Investig. Ophthalmol. Vis. Sci.* **2015**, *56*, 3790.
44. Kuha, J. AIC and BIC: Comparisons of Assumptions and Performance. *Sociol. Methods Res.* **2004**, *33*, 188–229. [CrossRef]
45. Wei, J.; Zhou, L. Model selection using modified AIC and BIC in joint modeling of paired functional data. *Stat. Probab. Lett.* **2010**, *80*, 1918–1924. [CrossRef]
46. Galecki, A.; Burzykowski, T. *Linear Mixed-Effects Models Using R*; Springer: New York, NY, USA, 2013.
47. Kutner, M.; Nachtsheim, C.; Neter, J.; Li, W. Chapter 2: Inference in Regression and Correlation Analysis. In *Applied Linear Statistical Models*; McGraw-Hill/Irwin: New York, NY, USA, 2005; pp. 63–73.
48. Fernandez, G.C.J. Residual Analysis and Data Transformations: Important Tools in Statistical Analysis. *HortScience* **1992**, *27*, 297–300. [CrossRef]
49. Cook, R.D.; Weisberg, S. *Residuals and Inlfuence in Regression (Monographs on Statistics and Applied Probability)*; Chapman and Hall: New York, NY, USA, 1982.
50. Bowerman, B.L.; O'Connell, R.T.; Koehler, A.B. *Forecasting, Time Series, and Regression*; Brooks/Cole, a Part of Cengage Learning; Thomson Brooks/Cole: Boston, MA, USA, 2005.
51. Brandimarte, P. *Introduction to Distribution Logistics*; Wiley: Hoboken, NJ, USA, 2007.

52. Skinner, D.R.; Benke, K.K.; Chung, M.J. Application of adaptive convolution masking to the automation of visual inspection. *IEEE Trans. Robot. Autom.* **1990**, *6*, 123–127. [CrossRef]
53. Conger, A.J. A Revised Definition for Suppressor Variables: A Guide to Their Identification and Interpretation. *Educ. Psychol. Meas.* **1974**, *34*, 35–46. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.