Automatic Explainable Progress Prediction of Wet Age-Related Macular Degeneration

Riccardo Fidanza[1 [0009−0008−6604−7991]], Daniele Meli[1 [0000−0002−3162−388X]], and Emilia Maggio[2 [0000−0001−5656−820X]]

[1] Department of Computer Science, University of Verona, Italy
[2] Ophthalmology Unit, IRCCS Sacro Cuore Don Calabria Hospital Verona, Italy

Abstract. Wet age-related macular degeneration (wAMD) is an aggressive pathology representing a leading cause of central vision loss in the elderly population. Currently, diagnosis and treatment evaluation rely on clinical experts visually interpreting tomography scans to identify pathological features for decision-making, a process that is time-consuming and subject to inter-observer variability.
Stemming from the success of artificial intelligence (AI) in medicine, this paper proposes two main contributions. First, we address the **novel task of automatic diagnosis and monitoring of the treatment outcome for wAMD**. We adopt a **unique and novel dataset** containing recordings from 275 patients over different years of treatment. We show that different **AI models significantly outperform the recall (measuring misclassified wAMD worsening) of human evaluation (+20% at least)**. As a second contribution, we perform an **explainability study** on the trained AI models, evidencing that the **relevant features guiding the predictions are indeed a smaller subset and clinically relevant**. Our results pave the way towards **trustable automatic diagnosis and treatment evaluation for wAMD and related pathologies**, reducing significantly the effort required from clinicians.

Keywords: Explainable AI · Ophthalmology · Computer-assisted medicine

1 Introduction

Age-related macular degeneration (AMD), and especially its wet version (wAMD), is one of the leading causes of irreversible central vision loss in older adults, especially in developed countries [3]. Despite advances in treatments (notably anti-VEGF drug injections that can slow disease progression), wAMD remains a major cause of severe vision impairment. Clinicians typically analyze optical coherence tomography (OCT) scans of the eyes of the patients, in order to identify relevant features which can support the diagnosis and the evaluation of treatment response [2]. However, this process is time-consuming, error-prone, and requires a significant effort [5]. Artificial intelligence (AI) has shown promising

results in the detection of different types of wAMD [6]. However, one fundamental unanswered question is whether AI can also support the evaluation of treatment outcome, reducing the burden for the clinician.

This paper provides two contributions. First, we consider the **novel task of automatic prediction of wAMD progression after treatment**. To this aim, we use a **unique and novel dataset** collected at IRCCS Sacro Cuore Don Calabria Hospital in Verona (Italy), consisting of the historical treatment data of 235 patients across multiple years of monitoring, annotated with 90 standard clinical features. We then compare different AI models at predicting the status of wAMD, also in view of the standard performance by clinicians in literature. Secondly, we employ *explainable AI (XAI)* techniques to highlight relevant features for the trained models, and **analyze both the clinical importance and the contribution of these features to the classification performance**.

2 Novel wAMD dataset

Our novel dataset[3] consists of 325 observations, each representing a single eye affected by AMD. These observations were collected from 275 patients under treatment for wAMD monitored at IRCCS Sacro Cuore Don Calabria Hospital in Verona (Italy), with some patients contributing data from both eyes. Each sample is described by 90 features annotated by clinicians, providing a comprehensive and heterogeneous profile of both the eye and the patient's overall medical context. Among the 90 available numerical features, notable ones include the duration of the follow-up period (FU); the best corrected visual acuity (BCVA), measured at various time points during treatment; the ellipsoid zone (EZ), external limiting membrane (ELM) and thickness (CT) of the retina in the macular area, the intraretinal (IRF) and subretinal fluids (SRF); the neovascular lesion area (NLA).

3 Methodology

In this section, we present the main methods for explainable wAMD progress classification.

3.1 Data Preprocessing

To ensure data quality and usability, the dataset underwent a thorough cleaning process. First, the presence of missing values (NaNs) was carefully examined across all features. Features with a high percentage of missing data (greater than 10%) were removed entirely to avoid introducing bias or noise, reducing the total number of features from 90 to 63. For features with a low percentage of missing values, imputation was performed using the mode (i.e., the most frequent value) of each respective feature, given the mostly categorical / discrete nature of most

[3] **The dataset was collected with patients' consent and anonymized.**

features. Finally, categorical variables were encoded into numerical representations to enable the application of machine learning classification algorithms.

3.2 AI models

Different AI models were selected to leverage different strengths in handling the dataset's characteristics, such as feature heterogeneity, nonlinearity and interpretability. Hyperparameter tuning was performed via grid search and cross validation based on the accuracy metric.

We selected the following AI models as the most popular in the body of literature in other ophthalmologic classification tasks [1, 6]: **decision tree (DT)** and **Random Forest (RF)**, tree-based models chosen for their interpretability and ability to handle nonlinear relationships, categorical data and numerical data without normalization; **linear and nonlinear support vector classifier (SVC)**, selected for their efficiency and performance at class separation; **logistic regression (LR)** for interpretability as a baseline model; an **artificial neural network (ANN)** consisting of 2 fully connected layers with ReLU activations to capture complex inter-feature dependencies, and a dropout rate of 0.5 to mitigate overfitting.

3.3 Explainability analysis

To identify the most informative features and reduce dimensionality, recursive feature elimination with cross-validation (RFECV) was employed. This method recursively removes the least important features based on feature importance scores provided by an estimator and evaluates model performance at each step using cross-validation. In this case, a Random Forest classifier was used as the underlying estimator due to its ability to provide reliable and interpretable feature importance rankings. RFECV automatically selects the optimal number of features by maximizing the cross-validated accuracy score. This process not only enhances the generalization of the model, but also offers valuable insights into which features contribute the most to the classification task.

4 Classification results

We trained the classification models with 5-fold cross-validation, and evaluated them according to precision, recall, F1-score and accuracy. Since the dataset is slightly imbalanced, with approximately 64.3% negative (i.e., worsened wAMD) points vs. 35.7% positive entries (i.e., improved wAMD), we report median and IQR, and also evaluate ROC curves.

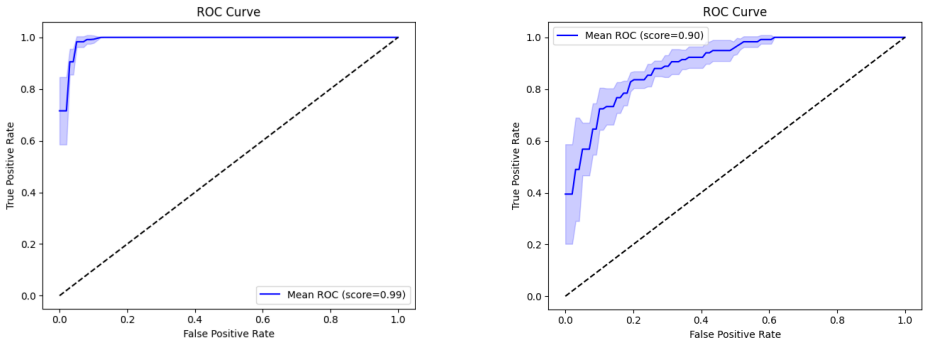
Table 1 and Figure 1a show the results obtained using the full dataset of 63 features (after data cleaning).

All models demonstrate good performance in the classification task, with comparable metrics and especially **very high recall with respect to known clinical performance [4] ($> 89\%$ vs. 62%). The recall identifies the number of**

Table 1: Classification performance with the full 63 features (Table 2 left).

Model	Accuracy		Precision		Recall		F1 Score	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Decision Tree	0.9231	0.0154	0.9242	0.0050	0.9231	0.0154	0.9234	0.0168
Random Forest	**0.9692**	**0.0154**	**0.9692**	**0.0159**	**0.9692**	**0.0154**	**0.9689**	**0.0152**
Nonlinear SVM	0.8923	0.0154	0.8979	0.0141	0.8923	0.0154	0.8927	0.0159
Logistic Regression	0.9231	0.0308	0.9290	0.0315	0.9231	0.0308	0.9240	0.0311
ANN	0.9385	0.0615	0.9286	0.0581	0.9512	0.0584	0.9359	0.0602

misclassified wAMD worsening instances. In particular, the RF classifier demonstrates a clear distinction by attaining the highest median accuracy and the lowest variability across metrics, thereby evidencing its superior robustness and reliability.



(a) Full 63 features

(b) RFECV 24 features

Fig. 1: ROC curves.

5 Explainability results

We apply RFECV to the best-performing RF classifier, in order to highlight the relevant features behind its high scores and gain interesting clinical outcomes.

Table 2 (**left**) shows the top-5% important features. **Two clinical experts agree on the well-known relevance of these features in the ophthalmologic literature, evidencing the trustworthy and reliable decision-making process behind the AI model**.

In order to gain deeper clinical insights, Table 2 (**right**) reports the results of RFECV (top-5%) when applied to the RF classifier, but *excluding the well-known relevant features reported on the left of Table 2*. **The clinicians also**

Table 2: Selected features from the full dataset with their importances.

Feature	Importance
EZ last visit	0.2456
BCVA 8 years	0.2378
BCVA 7 years	0.1935
ELM last visit	0.1379
BCVA 6 years	0.0724
FU (years)	0.0596
BCVA 5 years	0.0533
BCVA 4 years	0.1978
NLA	0.1361
BCVA 3 years	0.1297
BCVA 6 years	0.0905
BCVA 2 years	0.0755
Final CT	0.0616
BCVA 1 years	0.0609
Predom Persistent IRF	0.0587

acknowledge the relevance of these features (24 total), though usually not considered in the available medical literature, hence representing an advancement in clinical knowledge.

Table 3: Classification performance with 24 RFECV features (Table 2. right).

Model	Accuracy		Precision		Recall		F1 Score	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Decision Tree	0.7692	0.0462	0.7870	0.0249	0.7692	0.0462	0.7703	0.0373
Random Forest	**0.8462**	**0.0308**	**0.8445**	**0.0352**	**0.8462**	**0.0308**	**0.8404**	**0.0368**
Nonlinear SVM	0.8154	0.0308	0.8251	0.0211	0.8154	0.0308	0.8051	0.0275
Logistic Regression	0.8154	0.0154	0.8271	0.0056	0.8154	0.0154	0.8182	0.0134
ANN	0.7077	0.0923	0.7833	0.1340	0.6005	0.1160	0.6038	0.1958

In order to further validate this finding, in Table 3 and Figure 1b we report the performance of the AI classifiers *considering only the newly discovered relevant features of Table 2 (right)*. Compared to the results obtained in Table 1, the performance decreases by approximately 10% over all metrics, remaining well above the state-of-the-art human capabilities (**RF achieves 85% vs. 62% recall [4]**). Moreover, the ANN results much more sensitive to the changes in the feature set, showing its lower robustness and interpretability.

Overall, our findings indicate that the **AI models rely on clinically-relevant features** (Table 2), and that **a limited set of features (24/63) usually not reported in established ophthalmologic literature have an important impact on classification accuracy** (Table 3), still consistently outperforming humans.

6 Conclusion

This paper addresses the **novel and crucial problem of automatic wAMD treatment outcome evaluation with AI**. The results on **our unique dataset containing records of 235 patients** evidence that **AI models (especially**

Random Forests) significantly outperform human capacities (96% vs. 62% recall [4]). Moreover, our XAI study shows that the **AI model relies its predictions on clinically-relevant features**. More specifically, some of these features are well-known to the ophthalmologic community, evidencing the trustability of AI for this application. On the other hand, **a subset of other automatically extracted 24/63 features is recognized by the clinicians to advance the current medical knowledge**, while yielding slightly worse classification performance (**still well above the human performance, 85% vs. 62% recall**). **This potentially significantly shrinks the range of features to be annotated from OCT scans, relaxing the effort from the clinicians.**

This paper represents a preliminary, yet fundamental step towards the adoption of trustworthy AI for the diagnosis and monitoring of wAMD and related pathologies. In the future we plan to address several limitations of this work. First, we will enrich our unique dataset to solve some imbalance problem, and make it available to the research community. Moreover, we will investigate the problem of predicting the wAMD evolution after treatment starting directly from OCT scans, thus not requiring any dataset annotation.

References

1. Desai, R.J., Wang, S.V., Vaduganathan, M., Evers, T., Schneeweiss, S.: Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA network open **3**(1), e1918962–e1918962 (2020)
2. Hu, Y., Gao, Y., Gao, W., Luo, W., Yang, Z., Xiong, F., Chen, Z., Lin, Y., Xia, X., Yin, X., et al.: Amd-sd: An optical coherence tomography image dataset for wet amd lesions segmentation. Scientific Data **11**(1), 1014 (2024)
3. Kansagara, D., Gleitsmann, K., Gillingham, M., O'Neil, M., Saha, S.: Nutritional supplements for age-related macular degeneration: A systematic review. Tech. rep., Department of Veterans Affairs (US), Washington (DC) (Jan 2012), https://www.ncbi.nlm.nih.gov/books/NBK84266/, vA Evidence-based Synthesis Program Reports
4. Liefers, B., Taylor, P., Alsaedi, A., Bailey, C., Balaskas, K., Dhingra, N., Egan, C.A., Rodrigues, F.G., Gonzalo, C.G., Heeren, T.F., et al.: Quantification of key retinal features in early and late age-related macular degeneration using deep learning. American Journal of Ophthalmology **226**, 1–12 (2021)
5. Mete, M., Iacovello, D., Maggio, E.: Novel diagnosis and therapeutics approaches in retina diseases (2024)
6. Wang, Z., Keane, P.A., Chiang, M., Cheung, C.Y., Wong, T.Y., Ting, D.S.W.: Artificial intelligence and deep learning in ophthalmology. In: Artificial intelligence in medicine, pp. 1519–1552. Springer (2022)