Machine Learning-Based Muscle Response Analysis for Diabetic Retinopathy: Trial Dynamics and Diagnostic Value

Machine Learning-Based Muscle Response Analysis for Diabetic Retinopathy: Trial Dynamics and Diagnostic Value

V. Negri, A. Laffi, M. Raffi, A. Piras, A. Mingotti, R. Tinarelli

***Abstract*—Diabetic retinopathy (DR) is a major cause of vision impairment worldwide, and early diagnosis remains a key challenge. Building upon previous work that demonstrated the utility of machine learning (ML) in classifying DR patients based on electromyographic (EMG) muscle response to visual stimuli, this study extends the investigation by analyzing inter-trial dynamics, test duration efficiency, and subject-aware cross-trial transferability. A Random Forest (RF) classifier was trained on muscle response data collected from healthy individuals, untreated DR patients, and laser-treated patients, each exposed to structured visual stimuli across multiple trials. The study evaluates classification performance under varying conditions: different acquisition lengths, reduced test durations, inter-trial transfer scenarios within the same subject, and in the presence of measurement uncertainty. Results confirm that patient class and stimulus type can be reliably predicted even with short-duration acquisitions, highlighting that early EMG signal segments encode sufficient discriminative information. Additionally, performance trends across trials provide insights into response stability and intra-subject adaptation effects over time. This work reinforces the potential of ML-assisted neuromuscular analysis in DR diagnosis within controlled, subject-aware measurement scenarios and suggests pathways for more efficient, personalized screening protocols.**

***Index Terms*—Diabetic Retinopathy, diagnostic optimization, electromyography, medical measurement, muscle responses, Random Forest, optic flow stimulation, trial variability.**

I. INTRODUCTION

DIABETES is a systemic disease marked by chronic dysregulation of blood glucose levels, which can progressively lead to vascular and neurological complications [1]. Among its most serious consequences is diabetic retinopathy (DR), a condition in which prolonged hyperglycemia damages the retina's microvasculature. While DR may initially be asymptomatic or cause only mild visual disturbances, its progression can result in microaneurysms, hemorrhages, and fluid leakage, ultimately leading to significant visual impairment or blindness. Effective glycemic and blood pressure control remains the primary preventive strategy, while laser photocoagulation is commonly used in more advanced stages to stabilize the condition by sealing leaking vessels and reducing neovascularization [2].

Beyond visual degradation, DR has systemic implications that are often underexplored. Vision is a key contributor to multisensory integration for postural control, providing real-time feedback to neural systems that govern balance and movement [3]. The retina serves as a critical interface between the external visual environment and the cerebellum, and its dysfunction, such as that caused by DR, can impair neuromuscular coordination. This disruption contributes to increased postural instability and fall risk in diabetic patients [4].

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) techniques have shown growing potential in enhancing the diagnosis and understanding of DR and its broader physiological effects. For example, Genetic Algorithms and Fuzzy C-Means have been used to improve angiographic image segmentation for DR prediction [5], while Random Forest (RF) and Logistic Regression models have helped identify systemic fall risk factors in diabetic populations [6]. Other works have modeled ocular muscle behavior to simulate disease dynamics [7], explored protein interaction networks to identify new DR-related genes [8], and integrated multiple ML methods for biomarker discovery and early detection [9]. Non-invasive approaches, such as Electrooculography (EOG) combined with ML classifiers, have also emerged as accessible diagnostic alternatives [10]. Authors in [11] propose a novel architecture, C2x-FNet, combining cascaded dense blocks with a twofold cross-feature enhancement module (CDB-2X-FEM) to improve diabetic retinopathy grading by capturing subtle abnormalities and enhancing contextual understanding, achieving superior performance on three benchmark datasets. The study [12] introduces KMoCoL, a two-stage k-positive momentum contrastive learning framework for image-based diabetic retinopathy grading. By generating balanced and semantically rich features, the method addresses class imbalance and achieves state-of-the-art performance on two benchmark fundus image datasets. Finally, [13] presents the HFRF-Net, a hierarchical full-resolution fusion network that combines spatial and contextual paths with novel attention and enhancement blocks to improve the accuracy and connectivity of retinal blood vessel segmentation, achieving superior results on three benchmark datasets. More broadly, the adaptability of ML to health-related signal processing is further demonstrated in [14], which applies advanced deep learning architectures to behavioral monitoring in Ambient Assisted Living contexts.

However, despite this growing body of research, very limited attention has been devoted to exploring neuromuscular responses as a diagnostic marker of DR. Building on this

TABLE I. PATIENT GROUP CHARACTERISTICS.

	Retinopathy	Laser	Control
Subjects (N°)	12 (6 M; 6 F)	8 (2 M; 6 F)	12 (8 M; 4 F)
Age (years)	62 ± 3	58 ± 5	58 ± 2
BMI (kg/m^2)	28 ± 4	28 ± 3	26 ± 4
Age at diabetes onset (years)	37 ± 4	30 ± 5	-
Disease duration (years)	25 ± 3	29 ± 2	-
HbA_{1c} (%)	8.1 ± 1.2	8.4 ± 0.9	-

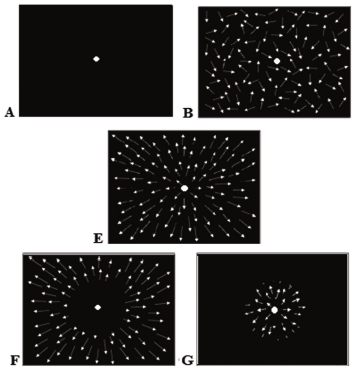

Fig. 1. Optic Flow Stimuli: Baseline (A), Random (B), DC-FC (E), Periphery (F), and Fovea (G).

landscape, [15] introduced a novel ML-based method for DR assessment using electromyographic (EMG) muscle response data from subjects exposed to visual stimuli. An RF classifier was trained to differentiate between healthy individuals, DR patients, and laser-treated patients and to infer the administered visual stimulus. The study demonstrated that DR could leave subtle traces in neuromuscular activation, detectable by ML even when invisible to human inspection. Nevertheless, its scope was limited to feasibility testing under controlled conditions and did not address several aspects that are critical for the practical use of EMG-based analysis in DR, particularly in repeated and subject-dependent measurement scenarios. This work, a technical extension of [15], introduces a structured, subject-aware evaluation framework that provides new experimental evidence along the following directions:

- **Temporal efficiency of EMG-based discrimination:** Classification performance is analyzed as a function of EMG acquisition duration, from full-length recordings down to ≤1 s windows. The results show that the information exploited by the model is concentrated in the early phase of the muscle response, providing quantitative guidance for the design of measurement protocols.
- **Intra-subject consistency across repeated trials:** Model performance is analyzed across multiple trials acquired from the same subject within a single session, providing evidence on response repeatability under repeated visual stimulation and revealing intra-subject variability relevant for personalized assessment.
- **Cross-trial transferability within subject-aware settings:** Inter-trial transfer scenarios are considered to assess whether models trained on a given acquisition remain valid across subsequent trials of the same subject, extending beyond conventional within-trial validation and supporting model reuse in repeated-measurement contexts.
- **Robustness under realistic measurement uncertainty:** Framework reliability is evaluated by introducing controlled perturbations into the EMG signals and repeating training and testing over multiple realizations, characterizing performance stability under noisy and imperfect acquisition conditions.

Through these contributions, the present study further validates the diagnostic relevance of EMG-based muscle response patterns in DR by providing quantitative evidence on temporal efficiency, intra-subject consistency, cross-trial transferability, and robustness to uncertainty. The results also confirm that brief, non-invasive EMG recordings are sufficient for accurate prediction of both patient condition and stimulus type within controlled, subject-aware acquisition scenarios.

The remainder of the paper is organized as follows. Section II describes the data acquisition process, subject groups, and preprocessing steps. Section III outlines the ML framework and experimental design. Section IV presents the results across all evaluation tests, and Section V discusses implications, limitations, and future directions. Conclusions are drawn in Section VI.

II. MATERIALS AND METHODS

A. Data Acquisition

Thirty-two participants gave written informed consent and were divided into three groups: 12 with early-stage DR, 8 who had undergone peripheral laser treatment (eccentricity > 30°), and 12 healthy controls. Table I summarizes sex distribution and group averages (± SD) for age, BMI, diabetes onset age, disease duration, and HbA1c. Exclusion criteria included musculoskeletal disorders, CNS-affecting medications, and diabetes complications beyond retinopathy that might affect posture (see [4]).

Participants stood barefoot in a quiet stance inside a dark room, facing a large screen displaying white dots as optic flow stimuli (5°/s). They were instructed to fixate centrally throughout. Figure 1 shows the five stimulus types: Baseline (dark screen), Random, DC-FC (full field), Periphery (excluding central 20°), and Fovea (7° radius), each shown in five 30-second trials. Postural sway was recorded via two Kistler® force platforms, capturing COP displacement in anteroposterior and mediolateral directions. EMG signals were recorded at 1 kHz using BTS PocketEMG, with electrodes on the tibialis anterior and soleus muscles per SENIAM guidelines (www.seniam.org). EMG signals were band-pass filtered (20-450 Hz) and normalized to maximum voluntary contraction for cross-subject comparison.

Each participant received both written and verbal explanations of the experimental procedures, and informed consent was obtained prior to data collection. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Bioethics Committee of the University of Bologna (protocol code: 0325016, approval date: December

18, 2019).

B. Data Preprocessing

To prepare the data for the ML model, a comprehensive dataset combining EMG recordings and patient characteristics was created. Data from three patient groups – healthy, retinopathic, and laser-treated – were included, with each subject exposed to five different visual stimuli. Each subject was assigned a unique identifier to enable the organization of repeated measurements across trials. All continuous features, including acquisition time and EMG muscle response signals, were normalized to the [0,1] range to ensure comparability across subjects and sessions. Visual stimulus types were label-encoded, while patient identifiers were one-hot encoded to allow their inclusion as categorical variables representing subject-specific information.

Depending on the classification task, the input–output configuration was defined as follows. For patient classification, the input feature set consisted of normalized EMG signals, acquisition timestamps, one-hot encoded patient identifiers, and label-encoded stimulus types, while the output corresponded to the patient category (healthy, retinopathic, or laser-treated). For stimulus classification, the same input features were used, except that the stimulus type served as the output variable and was therefore excluded from the input set.

Patient identifiers were included as subject-specific covariates to account for inter-subject variability across repeated measurements within the same experimental session. This choice reflects a controlled, subject-aware measurement scenario, in which patient identity is known (e.g., repeated or longitudinal assessments), rather than an anonymous population-level screening setting. Accordingly, patient IDs are used to model subject-dependent response characteristics and do not replace physiological discrimination based on EMG signals.

The final dataset included 12 healthy subjects, 12 retinopathic patients, and 8 laser-treated patients, with 30,000 frames per trial per stimulus, resulting in a total of 150,000 frames per subject per stimulus across the five stimulations.

C. Random Forest Algorithm

RF is a widely used machine learning algorithm known for its robustness and flexibility [16]. In this study, RF was selected as a baseline model due to its suitability for handling heterogeneous feature sets that combine continuous EMG signals and categorical variables, while requiring limited feature engineering and offering stable performance in the presence of noise and repeated measurements. This makes RF particularly appropriate for an exploratory, measurement-oriented analysis aimed at assessing the feasibility of ML-based neuromuscular analysis for DR, rather than optimizing performance through complex model architectures. The RF model was implemented using the Scikit-learn library [17]. Based on preliminary prior results of [15], the number of trees was set to 10 for patient classification and 50 for stimulus classification, as these values ensured stable classification performance while maintaining low computational cost. Parallel processing was employed to reduce training time.

Model training and testing were performed using an 80/20 split of the available data. This choice was validated by earlier 20-fold cross-validation experiments reported in [15] and was adopted consistently across all experiments in the present study to ensure clarity, reproducibility, and comparability of results. All reported performances should therefore be interpreted within the defined subject-aware classification framework, consistent with the experimental protocol and the intended application to repeated or longitudinal measurements.

III. Experimental Design

A. Evaluation Approach

The evaluation focused on two distinct classification tasks: patient classification and stimulus classification. For patient classification, the goal was to predict the category of each participant as healthy, diabetic retinopathic, or laser treated. For stimulus classification, the model predicted the type of visual stimulus based solely on EMG activity, testing the hypothesis that neuromuscular responses encode distinct patterns corresponding to the presented stimuli. Accurate classification of visual stimuli would indicate a systematic relationship between visual input and postural adjustments, providing new insights into sensorimotor integration in DR.

Two datasets were considered: one limited to responses from two specific stimuli (baseline and DC-FC), and another that included all five stimulus types (described in subsection II.A).

In addition to varying the input features and the considered stimuli, the following tests were performed:

a) **Test 1: Classification vs. Number of Samples**: Various data acquisition lengths were evaluated to assess the impact on classification performance. The data used was from the first trial, with 15,000, 5,000, and 1,000 samples per acquisition file, corresponding to 15, 5, and 1 s of acquisition. Additionally, for the 1,000 samples scenario, the data was progressively split into steps, with the model tested on smaller chunks. The output of the model was either the patient type or the applied stimulus type, considering both 2-stimulus and 5-stimulus datasets.

b) **Test 2: Performance vs. Trials (1 kSa, 5 Stimuli)**: This test assessed the model's performance by varying the trial used, specifically using data from Trials 1 and 5, with 1,000 samples per acquisition file. The output was either the class or the stimulus, considering the 5-stimulus dataset. The analyses consider two main scenarios: one with a single patient per class and another with three patients per class. This approach was chosen to investigate the effect of inter-patient variability on model performance while maintaining controlled conditions to identify emerging patterns. The decision to reduce the number of considered patients was necessary because some subjects missed complete data for all five trials, due to the exclusion of sessions with inconsistent or incomplete recordings.

c) **Test 3: Cross-Trial Testing (1 kSa, 5 Stimuli)**: This test evaluated the model's cross-trial generalization by training on Trial 1 and testing on Trials 2 to 5, using a 1,000-sample input. The output was either the class or the stimulus, considering the 5-stimulus dataset. The

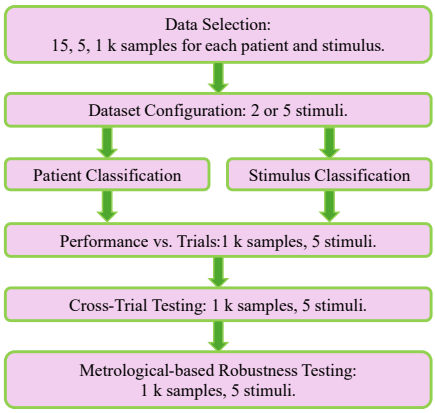

Fig. 2. Flowchart of the RF Evaluation Approach.

TABLE II. CLASSIFICATION REPORT OF PATIENT CLASS CLASSIFICATION – 15,000 SAMPLES – 2 STIMULI DATASET.

	Precision	**Recall**	**F1**	**Support**
Healthy	100 %	100 %	100 %	71960
Retinopathic	100 %	100 %	100 %	71960
Laser-treated	100 %	100 %	100 %	48065
Macro-Avg	100 %	100 %	100 %	192000
Accuracy	100 %			192000

experiment was executed with both 1 and 3 patients for class, for the reasons previously discussed.

d) **Test 4: Metrological-based Robustness Testing (1 kSa, 5 Stimuli)**: This test evaluated the model's robustness by training and testing with corrupted data, using 1,000 samples for each patient and stimulus from the 5-stimuli dataset. Particularly, different perturbation levels have been introduced within the data, to model the measurement uncertainty. Each EMG signal point has been corrupted with random noise from a uniform distribution with bounds determined by the associated uncertainty. To comprehensively evaluate the consistency of results, 1, 2, 5, and 10 % perturbation levels have been selected to use a broader range of typical uncertainty associated with medical data [18]. Moreover, to assess the repeatability of the results, data corruption and model training and testing have been repeated 100 times, across all levels of uncertainty.

Figure 2 shows the flow diagram of the dataset configurations and the performed test, highlighting the key aspects of the evaluation approach.

B. Evaluation Metrics

In evaluating an ML classifier for a multi-class classification task, several key metrics can be used to assess performance [19]. In this study, Accuracy, Precision, Recall, and F1-score have been selected. Furthermore, confusion matrices are reported.

Accuracy measures the overall correctness of the model by computing the ratio of correctly classified instances to the total number of instances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

where True Positives (TP) are instances correctly classified as belonging to a given class, True Negatives (TN) are instances correctly classified as not belonging to a given class, False Positives (FP) are instances incorrectly classified as belonging to a given class when they belong to another class, and False Negatives (FN) are instances that actually belong to a given class but are incorrectly classified as another class. been selected.

Precision evaluates how many of the instances predicted as a specific class are correct among all the predictions made for that class:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

Recall indicates how many instances of a class are correctly identified by the model among all the actual instances of that class:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

F1-score is the harmonic mean of Precision and Recall, providing a balanced measure when there is an uneven class distribution:

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

Since this is a multi-class classification problem, also the Macro-Average (Macro-Avg) is used to compute the average of Precision, Recall, and F1-score across all classes.

Finally, the Confusion Matrix presents a detailed report of model predictions, showing TP, FP, TN, and FN for each class.

III. EXPERIMENTAL RESULTS

A. Test 1: Classification vs. Number of Samples

Firstly, the results of the patient classification task are reported. Table II reports the classification performance of the RF algorithm applied to patient classification, using muscle response data from two stimuli (Baseline and DC-FC), with 15,000 samples selected per patient and stimulus. Note that the Support column indicates the number of test instances for each class, providing context on the sample size used to evaluate the corresponding performance metrics. The results indicate perfect model performance, with 100 % Precision, Recall, and F1-score across all classes, as well as an overall Accuracy of 100 %.

Figure 3 presents the confusion matrix for patient classification using the dataset with five stimuli. For each patient and stimulus, 15,000 samples were used. Despite the larger and more complex dataset, which includes data collected from patients subjected to five different visual stimuli, only two patients are incorrectly identified out of

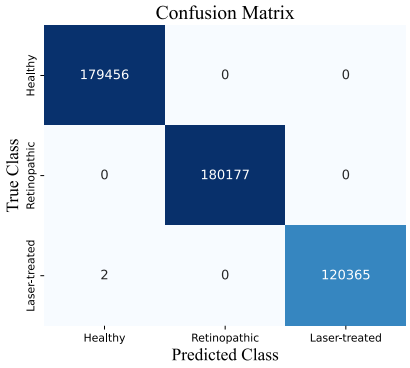

Fig. 3. RF Confusion Matrix of Patient Class Classification – 15,000 Samples – 5 Stimuli Dataset.

TABLE III. CLASSIFICATION REPORT OF STIMULUS TYPE CLASSIFICATION – 15,000 SAMPLES – 2 STIMULI DATASET.

	Precision	Recall	F1	Support
Baseline	97 %	98 %	98 %	96211
DC-FC	98 %	97 %	98 %	95789
Macro-Avg	98 %	98 %	98 %	192000
Accuracy	98 %			192000

TABLE IV. CLASSIFICATION REPORT OF STIMULUS TYPE CLASSIFICATION – 15,000 SAMPLES – 5 STIMULI DATASET.

	Precision	Recall	F1	Support
Baseline	93 %	95 %	94 %	95975
DC-FC	95 %	93 %	94 %	95709
Fovea	94 %	94 %	94 %	95953
Periphery	94 %	93 %	93 %	95995
Random	94 %	95 %	94 %	96368
Macro-Avg	94 %	94 %	94 %	480000
Accuracy	94 %			480000

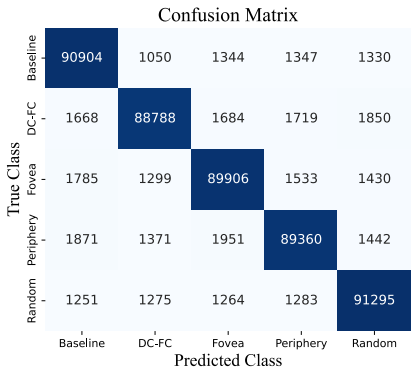

Fig. 4. RF Confusion Matrix of Stimulus Type Classification – 15,000 Samples – 5 Stimuli Dataset.

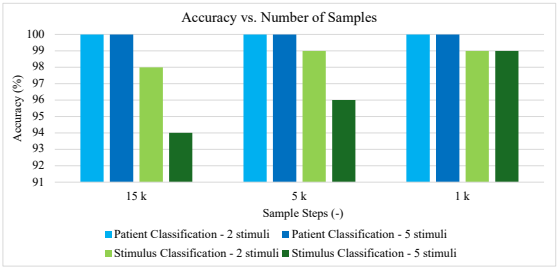

Fig. 5. RF Accuracy of Patient and Stimulus Type Classification vs. Number of Samples.

480,000 test instances, and the model maintains 100 % performance across all metrics.

Secondly, stimulus type classification task results are reported. Table III reports the classification performance of the RF algorithm applied to stimulus classification, using muscle response data from two stimuli (Baseline and DC-FC), with 15,000 samples selected per patient and stimulus. The results support the hypothesis that muscle response patterns encode information about the visual stimulus. With 98 % accuracy and Precision, Recall, and F1-score consistently at 97 – 98 %, the model effectively distinguishes between Baseline and DC-FC stimuli. The slight differences in Recall (higher for Baseline) and Precision (higher for DC-FC) indicate minimal misclassification.

Table IV and Figure 4 respectively present the classification report and confusion matrix for stimulus classification using the five-stimulus dataset and 15,000 samples for each patient and stimulus. The results demonstrate that, despite the increased complexity, the model still performs well with 94 % accuracy and Precision, Recall, and F1-scores around 93 – 95 % across all stimulus types. These results further support the hypothesis that there is a strong correlation between the type of visual stimulus and muscle response patterns, providing valuable insights into the physiological effects of DR.

To evaluate the effect of the considered number of samples per patient and stimulus type, Fig. 5 shows the accuracy values achieved in patient and stimulus type classification, using 15, 5, and 1 kSa for both the 2 and 5 stimuli datasets. Notably, the accuracy in patient classification is constant at the value of 100 %. On the other hand, when increasing the sample number to 1,000, stimulus classification tasks reach 99 % of accuracy for both 2 and 5-stimulus datasets.

This supports the idea that reducing the acquisition period to 5 or only 1 second (5,000 and 1,000 samples, respectively) does not compromise model performance. These findings underscore the influence of sample size on classification accuracy, demonstrating that fewer samples still yield highly accurate results, making data collection more efficient without sacrificing performance. Based on this result, the following tests are conducted using 1 kSa for each patient and stimulus type, and the 5-stimulus dataset, which represents the most complex task.

To further investigate the relation between the considered fragment of acquisitions and classifier performance, Figure 6 reports the accuracy of the RF model applied to stimulus type classification, using steps of samples. The analysis has been

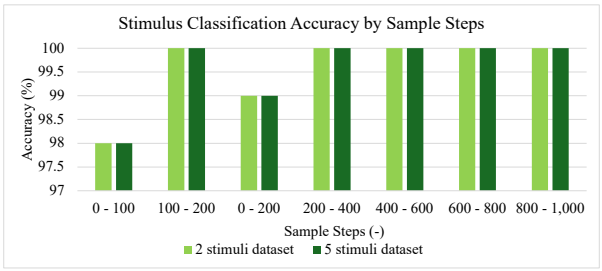

Fig. 6. RF Accuracy of Stimulus Type Classification – 1,000 Samples by Steps.

TABLE V. CLASSIFICATION REPORT OF PATIENT CLASSIFICATION – 1 PATIENT FOR CLASS – 1,000 SAMPLES – 5 STIMULI DATASET.

Trial	Class	Precision	Recall	F1	Support
1	Healthy	100 %	100 %	100 %	1052
	Retinopathic	100 %	100 %	100 %	989
	Laser-treated	100 %	100 %	100 %	959
	Macro-Avg	100 %	100 %	100 %	3000
	Accuracy	100 %			3000
5	Healthy	100 %	100 %	100 %	1019
	Retinopathic	100 %	100 %	100 %	986
	Laser-treated	100 %	100 %	100 %	995
	Macro-Avg	100 %	100 %	100 %	3000
	Accuracy	100 %			3000

TABLE VI. CLASSIFICATION REPORT OF STIMULUS TYPE CLASSIFICATION – 1 PATIENT FOR CLASS – 1,000 SAMPLES – 5 STIMULI DATASET.

Trial	Class	Precision	Recall	F1	Support
1	Baseline	99 %	100 %	99 %	598
	DC-FC	100 %	99 %	99 %	613
	Fovea	99 %	98 %	98 %	594
	Periphery	98 %	98 %	98 %	586
	Random	99 %	100 %	99 %	609
	Macro-Avg	99 %	99 %	99 %	3000
	Accuracy	99 %			3000
5	Baseline	100 %	100 %	100 %	620
	DC-FC	99 %	100 %	99 %	609
	Fovea	99 %	100 %	99 %	585
	Periphery	99 %	99 %	99 %	576
	Random	100 %	99 %	99 %	610
	Macro-Avg	99 %	99 %	99 %	3000
	Accuracy	99 %			3000

TABLE VII. CLASSIFICATION REPORT OF PATIENT CLASSIFICATION – 3 PATIENT FOR CLASS – 1,000 SAMPLES – 5 STIMULI DATASET.

Trial	Class	Precision	Recall	F1	Support
1	Healthy	100 %	100 %	100 %	2994
	Retinopathic	100 %	100 %	100 %	3019
	Laser-treated	100 %	100 %	100 %	2987
	Macro-Avg	100 %	100 %	100 %	9000
	Accuracy	100 %			9000
5	Healthy	100 %	100 %	100 %	3017
	Retinopathic	100 %	100 %	100 %	2957
	Laser-treated	100 %	100 %	100 %	3026
	Macro-Avg	100 %	100 %	100 %	9000
	Accuracy	100 %			9000

TABLE VIII. CLASSIFICATION REPORT OF STIMULUS TYPE CLASSIFICATION – 3 PATIENT FOR CLASS – 1,000 SAMPLES – 5 STIMULI DATASET.

Trial	Class	Precision	Recall	F1	Support
1	Baseline	99 %	98 %	98 %	1818
	DC-FC	99 %	99 %	99 %	1788
	Fovea	99 %	99 %	99 %	1769
	Periphery	99 %	99 %	99 %	1820
	Random	99 %	99 %	99 %	1805
	Macro-Avg	98 %	99 %	99 %	9000
	Accuracy	99 %			9000
5	Baseline	99 %	98 %	99 %	1739
	DC-FC	99 %	99 %	99 %	1890
	Fovea	99 %	99 %	99 %	1762
	Periphery	98 %	99 %	98 %	1815
	Random	99 %	99 %	99 %	1794
	Macro-Avg	99 %	99 %	99 %	9000
	Accuracy	99 %			9000

limited to stimulus classification, since patient classification performance remains constant to 100 % of accuracy. Results show that the muscular response that provides insights for the ML-based classification is contained in the first second of acquisition. As a matter of fact, steps from 100 to 1,000 samples reach a constant accuracy of 100 %.

B. Test 2: Performance vs. Trials (1 kSa, 5 stimuli)

The following results evaluate the performance of the RF model for classifying patient types and applied stimuli, based on data collected during trials 1 and 5, with 1,000 samples for each patient and stimulus, along with the 5-stimulus dataset.

Tables V and VI report the classification reports of patient and stimulus type classification using data with one patient per class, respectively. Patient classification achieved perfect performance across all metrics. Regarding stimulus classification, considering data from trials 1 and 5 introduces a slight variability in identifying the type of stimulus applied during the measurement campaign. However, performance remains around 98 – 100 %, with an overall accuracy of 99 %, demonstrating that the muscle response signals remain comparable during trial acquisitions and the number of test repetition do not affect patients' behavior.

This analysis is then extended to three patients for class. Tables VII and VIII report the results for patient and stimulus

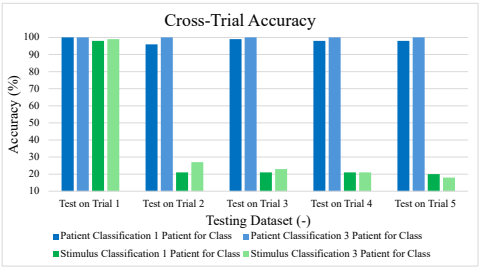

Fig. 7. RF Accuracy of Patient and Stimulus Type Classification in Cross-Trial Testing – 1,000 Samples – 5 Stimuli Dataset.

TABLE IX. ACCURACY MEAN AND STD. DEV. – 1,000 SAMPLES – 5 STIMULI DATASET – 100 REPETITIONS.

Output Type	Noise (%)	Accuracy (%)
Patient Class	1	100
	2	100
	5	100
	10	100
Stimulus Type	1	97
	2	97
	5	95
	10	92

TABLE X. ACCURACY MEAN AND STD. DEV. – 1,000 SAMPLES – 5 STIMULI DATASET – 100 REPETITIONS.

Output Type	Noise (%)	Mean (%)	Std. Dev. (%)
Patient Class	1	100.00	0.00
	2	100.00	0.00
	5	100.00	0.00
	10	100.00	0.00
Stimulus Type	1	100.00	0.04
	2	100.00	0.07
	5	99.21	0.36
	10	98.02	1.38

TABLE XI. TRAINING AND TESTING TIME FOR SOME CONSIDERED DATASETS.

Output Type	Dataset	Training time (s)	Testing time (s)
Patient Class	15,000 samples – 2 stimuli	2.03	0.09
	15,000 samples – 5 stimuli	5.69	0.05
	5,000 samples – 2 stimuli	0.56	0.02
	5,000 samples – 5 stimuli	1.72	0.04
	1,000 samples – 2 stimuli	0.13	0.01
	1,000 samples – 5 stimuli	0.25	0.02
Stimulus Type	15,000 samples – 2 stimuli	13.23	0.57
	15,000 samples – 5 stimuli	40.49	0.09
	5,000 samples – 2 stimuli	3.16	0.10
	5,000 samples – 5 stimuli	10.49	0.64
	1,000 samples – 2 stimuli	0.27	0.01
	1,000 samples – 5 stimuli	1.64	0.07

classification, respectively. These results confirm the slight performance variability in using data from trial 1 or 5.

C. Test 3: Cross-Trial Testing (1 kSa, 5 stimuli)

To further investigate the patient response dependency on test executions, cross-trial testing has been performed. Figure 7 shows the accuracy values for patient and stimulus type classification tasks with the RF model trained on data from trial 1 and tested on data from trials 1 to 5. The considered dataset includes 1,000 samples for each patient and stimulus type, considering all five stimuli.

Results demonstrated that for patient classification, performance remains near the maximum accuracy of 100 %, showing the generalization capability of the ML-based model across trials. On the other hand, for the stimulus type classification task, even if the performance vs. trials test shows a good consistency across all metrics, the response signals are not so similar to generalize between trials.

D. Test 4: Metrological-based Robustness Testing (1 kSa, 5 stimuli)

Table IX reports the accuracy values for patient and stimulus type classification tasks with different levels of applied noise, to assess the performance robustness in the presence of measurement uncertainty. Patient classification demonstrates stability, maintaining an accuracy of 100 % even in the case of corrupted data, across all the noise levels. On the other hand, stimulus classification is affected by the introduction of uncertainty, with an accuracy decreasing from 97 to 92 %.

To assess the repeatability of these results, Table X reports the mean and std. dev. of accuracy values, computed across the 100 repetitions of the model training and testing stages on randomly corrupted data. When increasing the number of repetitions, the models' performance remains more consistent even in the presence of a 10 % level of uncertainty.

V. DISCUSSION

The RF model demonstrated the ability to classify both patient condition and visual stimulus using EMG muscle responses, confirming that neuromuscular activation encodes information related to DR and visuomotor processing. Importantly, comparable performance was achieved using short acquisition windows (1-5 s), indicating that the discriminative information is concentrated in the early phase of the muscle response and enabling a substantial reduction in test duration.

Performance analysis across repeated trials showed high consistency for patient classification, while reduced cross-trial generalization was observed for stimulus classification, reflecting intra-subject variability in response patterns. This finding supports the adoption of subject-aware or personalized evaluation strategies in repeated-measurement scenarios.

For a comprehensive evaluation, Tab. XI reports the training and testing time for the RF algorithm in some of the considered cases. The results show that the proposed framework exhibited very limited training and testing times,

particularly for short acquisitions, and required minimal preprocessing, making it suitable for efficient experimental and clinical workflows.

Finally, a metrological-based evaluation of models' robustness has been performed to assess the consistency of results in the presence of measurement uncertainty. The findings demonstrate the high capability of ML models to classify both patient and stimulus type even under simulated noise, showing the applicability of the proposed approach in practical clinical applications.

Note that the reported 100 % accuracy in patient classification reflects the controlled, subject-aware evaluation framework adopted in this study, where patient identity is known and repeated measurements are available. While these results confirm the internal consistency of the approach under such conditions, subject-independent generalization to unseen patients remains a complementary and more challenging scenario to be addressed.

VI. Conclusion

This study demonstrated the potential of machine learning in enhancing diabetic retinopathy diagnosis through muscle response analysis. The Random Forest model effectively classified patient conditions and visual stimuli, showing that muscle responses could be a viable diagnostic tool for diabetic retinopathy within controlled, subject-aware measurement settings. By extending the initial work, this research addressed key factors such as temporal efficiency, trial dynamics, and evaluation regarding measurement uncertainty. The results indicated that accurate diagnoses could be achieved with short acquisition windows, as brief as one second, improving diagnostic efficiency. Additionally, the analysis of multiple trials revealed response variability, suggesting the need to account for intra-subject differences in clinical testing. The model's ability to maintain performance across trials and in the presence of measurement uncertainty highlighted its robustness, supporting its applicability in realistic experimental and clinical monitoring scenarios. Furthermore, the model performed well with increasing patient diversity, demonstrating its potential to handle inter-subject variability under the adopted evaluation framework. Overall, this work supports the feasibility of using muscle response data for diabetic retinopathy diagnosis, offering a faster, more efficient alternative to traditional methods for personalized evaluation contexts.

References

[1] Elliott T.L., Pfotenhauer K.M., Classification and Diagnosis of Diabetes, (2022) Primary Care - Clinics in Office Practice, 49 (2), pp. 191 - 200, DOI: 10.1016/j.pop.2021.11.011.

[2] A.H. Al Ghamdi, Clinical predictors of diabetic retinopathy progression; a systematic review, Curr. Diabetes Rev. 16 (2019) 242–247, https://doi.org/10.2174/1573399815666190215120435.

[3] M. Raffi, A. Piras, M. Persiani, M. Perazzolo, S. Squatrito, Angle of gaze and optic flow direction modulate body sway, J. Electromyogr. Kinesiol. 35 (2017) 61–68, https://doi.org/10.1016/j.jelekin.2017.05.008.

[4] Alessandro Piras, Monica Perazzolo, Sergio Zaccaria Scalinci, Milena Raffi, The effect of diabetic retinopathy on standing posture during optic flow stimulation, Gait & Posture, Volume 95, 2022, Pages 242-248, ISSN 0966-6362, https://doi.org/10.1016/j.gaitpost.2020.10.020.

[5] Ghoushchi S.J., Ranjbarzadeh R., Dadkhah A.H., Pourasad Y., Bendechache M., An Extended Approach to Predict Retinopathy in Diabetic Patients Using the Genetic Algorithm and Fuzzy C-Means, (2021) BioMed Research International, 2021, art. no. 5597222, DOI: 10.1155/2021/5597222.

[6] Suzuki Y., Suzuki H., Ishikawa T., Yamada Y., Yatoh S., Sugano Y., Iwasaki H., Sekiya M., Yahagi N., Hada Y., Shimano H., Exploratory analysis using machine learning of predictive factors for falls in type 2 diabetes, (2022) Scientific Reports, 12 (1), art. no. 11965, DOI: 10.1038/s41598-022-15224-4.

[7] Oliveira R.H.M., Silva M.S., Nunes G.A.M.A., Faria R.M., Santos K.S., Rosa L.L.F., Rosa M.F.F., Rosa S.S.R.F., Control engineering investigation of the effects of proliferative diabetic retinopathy on the crystalline lens and ciliary muscle dynamic behavior, (2023) Research on Biomedical Engineering, 39 (3), pp. 663 - 676, DOI: 10.1007/s42600-023-00297-5.

[8] Zhang J., Yang J., Huang T., Shu Y., Chen L., Identification of novel proliferative diabetic retinopathy related genes on protein–protein interaction network, (2016) Neurocomputing, 217, pp. 63 - 72, DOI: 10.1016/j.neucom.2015.09.136.

[9] Tao Y., Xiong M., Peng Y., Yao L., Zhu H., Zhou Q., Ouyang J., Machine learning-based identification and validation of immune-related biomarkers for early diagnosis and targeted therapy in diabetic retinopathy, (2025) Gene, 934, art. no. 149015, DOI: 10.1016/j.gene.2024.149015.

[10] Archana R., Rajalakshmi T., Vijay Sai P., Non-invasive technique to detect diabetic retinopathy based on Electrooculography signal using machine learning classifiers, (2022) Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 236 (6), pp. 882 - 895, DOI: 10.1177/09544119221085422.

[11] Madarapu S., Ari S., Mahapatra K., C2x-FNet: Cascaded Dense Block With Twofold Cross-Feature Enhancement Module for Diabetic Retinopathy Grading, (2025) IEEE Transactions on Instrumentation and Measurement, 74, art. no. 5000910, DOI: 10.1109/TIM.2024.3500044.

[12] Li L., Liu X., Hou X., Chen L., Zhou Y., Fu S., KMoCoL: k-Positive Momentum Contrastive Learning for Imbalanced Diabetic Retinopathy Grading, (2025) IEEE Transactions on Instrumentation and Measurement, 74, art. no. 5019012, DOI: 10.1109/TIM.2025.3542859.

[13] Su H., Gao L., Wang Z., Yu Y., Hong J., Gao Y., A Hierarchical Full-Resolution Fusion Network and Topology-Aware Connectivity Booster for Retinal Vessel Segmentation, (2024) IEEE Transactions on Instrumentation and Measurement, 73, art. no. 2521616, DOI: 10.1109/TIM.2024.3411133.

[14] S. Mari, F. Ciancetta, Á. Hernandez, D. Pizarro, L. de Diego-Otón and V. M. Navarro, "A Convolutional Transformer for Enhanced NILM in Human Activity Recognition," 2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Eindhoven, Netherlands, 2024, pp. 1-6, doi: 10.1109/MeMeA60663.2024.10596791.

[15] Negri V., Laffi A., Mingotti A., Tinarelli R., Raffi M., Piras A., "Enhancing Diabetic Retinopathy Diagnosis with Machine Learning: A Random Forest Approach Using Muscle Response Data," In Proceedings of the 2025 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Chania, Greece, May 28–30, 2025.

[16] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

[17] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[18] A. Chunovkina and A. Tumilovich, "Analysis and evaluation of data available in a medical laboratory at estimating measurement uncertainty," 021 XXXI International Scientific Symposium Metrology and Metrology Assurance (MMA), Sozopol, Bulgaria, 2021, pp. 1-4, doi: 10.1109/MMA52675.2021.9610980.

[19] El Mrabet, M.; El Makkaoui, K.; Faize, A. Supervised Machine Learning: A Survey. In Proceedings of the 2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet), Rabat, Morocco, 3–5 December 2021; pp. 1–10. https://doi.org/10.1109/CommNet52204.2021.9641998.