mathematics

MDPI

Article

Ranking Information Extracted from Uncertainty Quantification of the Prediction of a Deep Learning Model on Medical Time Series Data

Ruxandra Stoean [1], Catalin Stoean [1,*], Miguel Atencia [2], Roberto Rodríguez-Labrada [3] and Gonzalo Joya [4]

[1] Romanian Institute of Science and Technology, 400022 Cluj-Napoca, Romania; ruxandra.stoean@rist.ro
[2] Department of Applied Mathematics, Universidad de Málaga, 29071 Málaga, Spain; matencia@ctima.uma.es
[3] Cuban Neuroscience Center, 11600 Havana, Cuba; roberto.rodriguez@cneuro.edu.cu
[4] Department of Electronic Technology, Universidad de Málaga, 29071 Málaga, Spain; gjoya@uma.es
* Correspondence: catalin.stoean@rist.ro

Received: 1 June 2020; Accepted: 28 June 2020; Published: 2 July 2020

Abstract: Uncertainty quantification in deep learning models is especially important for the medical applications of this complex and successful type of neural architectures. One popular technique is Monte Carlo dropout that gives a sample output for a record, which can be measured statistically in terms of average probability and variance for each diagnostic class of the problem. The current paper puts forward a convolutional–long short-term memory network model with a Monte Carlo dropout layer for obtaining information regarding the model uncertainty for saccadic records of all patients. These are next used in assessing the uncertainty of the learning model at the higher level of sets of multiple records (i.e., registers) that are gathered for one patient case by the examining physician towards an accurate diagnosis. Means and standard deviations are additionally calculated for the Monte Carlo uncertainty estimates of groups of predictions. These serve as a new collection where a random forest model can perform both classification and ranking of variable importance. The approach is validated on a real-world problem of classifying electrooculography time series for an early detection of spinocerebellar ataxia 2 and reaches an accuracy of 88.59% in distinguishing between the three classes of patients.

Keywords: deep learning; time series; uncertainty quantification; Monte Carlo dropout; random forest

1. Introduction

Deep learning (DL) has uncovered substantial findings in supporting decision making in medicine. The high accuracy of its predictions has triggered its increasing use for a variety of clinical classification tasks [1]. However, in order to be truly effective in practice, the confidence in its output must be expressed beyond the resulting probabilities for the classes of the problem. Medical problems from the real-world have a limited number of available electronic records and often the data is imbalanced between classes. Moreover, other factors that derive from the experience of the clinician and influence the diagnosis are not encompassed within the data. This epistemic uncertainty together with the inherent aleatory unpredictability of the process can be tackled through methods for uncertainty quantification (UQ). The two best-known current practices in UQ are the use of Monte Carlo dropout (MCD) [2] at the prediction phase and the constitution of deep ensembles [3].

The current paper utilizes MCD to achieve UQ for the predictions of a deep architecture applied on a biomedical time series data. Uncertainty is quantified by performing a forward pass with a different dropout mask every time, and the result is a sample of outputs for the same input, instead of the typical single outcome. This process may be viewed as an ensemble of several instances of the same

network, but with distinct dropout masks. Uncertainty estimates will be subsequently reached from a statistical quantification (e.g., average probability, variance) of the obtained sample for each possible label of the problem. Data can be understood better analyzing the UQ results, as well.

A particular three-class problem of biomedical saccade classification from electrooculography was chosen as the application scenario in this UQ study for a twofold reason. There are various measurements (which are encoded as time series) made for each patient in turn and all these records are gathered in a register that is representative for the patient diagnosis. The diagnoses can be healthy (or control), presymptomatic and sick. However, not all records in a patient's register have a similar pattern, since a sick register can contain some records featuring a normal (control) profile, as well as there can be a few saccades that do not resemble healthy in one control set of samples. Naturally, the presymptomatic one will contain both control and sick samples.

Besides exhibiting a high degree of epistemic uncertainty, this medical task also requires the tailoring of the classical MCD to go beyond the UQ of the predictions for the individual records. Usually, the MCD is used to emphasize what are the samples where ambiguity occurs in classification and also for establishing a closer to optimal contour in segmentation problems. Herein we deal with a classification task based on time series, but the complexity of the problem arises from the fact that the data corresponding to the patient comprises a high number of samples. Accordingly, the UQ information is further used to establish the labeling per container (register of records). The manner of accumulating the uncertainty from the separate records and carrying it at the level of the registers represents a substantial contribution of the current study.

Accordingly, statistical measurements of mean and standard deviations are computed for each register from the MCD uncertainty estimates of average probability and variance of comprised saccades. The new features defining the registers are given to a Random Forest (RF) model that connects their values to the known diagnosis class.

The definition of registers through the statistical measurements of the UQ and their subsequent modelling by means of RF leads to a classification accuracy for the current test case scenario of 88.59%. The RF also provides the insight into which statistical variables of saccade composition direct the prediction towards a certain global label for the register.

The originality of the current approach stems from carrying UQ results from the level of individual saccades over to patient's registers, through the computation of means of averages and standard deviations from the estimates of probabilities and even variances that were obtained for the saccades. Furthermore, these are used to create a numerical data set that contains useful information for a Random Forest to establish accurate results at the level of the records. The research question that will be followed in the current article is whether the variance that was not used in the previous study [4] carries significant information to substantially increase the classification accuracy for the current data set.

The organization of the paper is as follows. Section 2 presents the complexity and particularities of the application problem, which gave rise to the theoretical generalization of the MCD procedure for the higher level of register classification. Section 3 outlines the novelty of the UQ framework with respect to the state of the art in biomedical signal modelling. The methodological flow is presented algorithmically and a visual exemplification is conducted on a selected register. Section 4 performs an analysis of the data based on what the UQ revealed. The experimental design and findings are given and discussed in Section 5. Finally, Section 6 summarizes the main conclusions of this work.

2. Materials

The problem investigated as the test case scenario for the generalization of Monte Carlo Dropout for Uncertainty Quantification in deep networks models regards the analysis of electrooculography time series for supporting an early diagnosis of spinocerebellar ataxia 2. This is a neurodegenerative disease that is incurable, but early action can be taken to slow its advancement once precursory signs are present and found.

The electrical potential of the eye of a person following the movement of an object on screen is recorded into electrooculograms. From these signals, saccades can be extracted, whose shape examination determines the physician to classify the patient into one of the three categories: control (C), presymptomatic (P) and sick (S). However, it does not suffice to inspect only one saccadic sample of a person to establish an accurate diagnosis, because the response of the same person at different times can take distinct saccadic shapes. In fact, one cannot even know a priori a universal number of samples of a certain shape that is sufficient to reach an outcome. Each register is thus made of a variable number of saccadic temporal vectors of several shapes and the diagnosis established by the physician at the end of this consultation. The complexity of the problem thus stems from the uncertainty regarding the exact register distribution of certain types of saccadic shapes that are strongly indicative of one of the three outcomes.

The data set has 85 registers (each corresponding to a different individual) and these contain 5953 saccades in total. Each saccadic sample has a length of 192. Three illustrative example registers, one pertaining to a class of the problem are shown in Figure 1 with their member saccades.

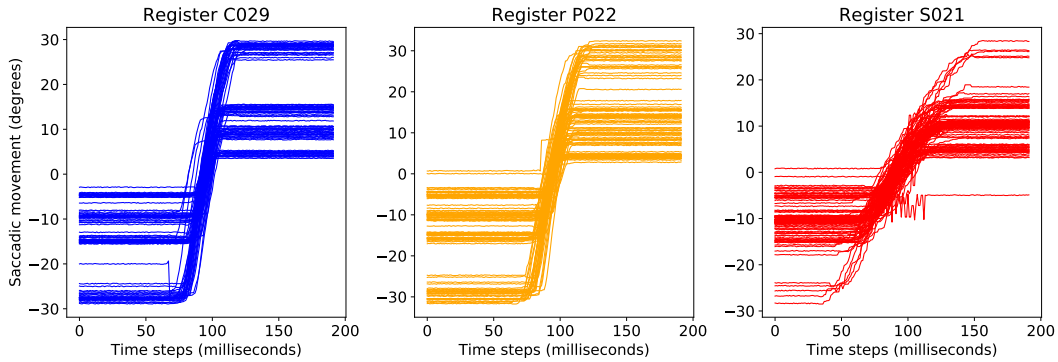

Figure 1. Representation of all saccades from a register in class control (C029), one presymptomatic (P022) and sick (S021).

This particular data set has been previously tackled by deep learning, where convolutional neural network (CNN) layers performed feature selection and a long short-term memory (LSTM) layer acquired the temporal pattern of the signal [5]. The clustering of the data before the DL application significantly increased the classification accuracy. A self organizing map provided a relabeling of the saccadic samples prior to learning that better connected the outcome with the shape similarity [6]. An early attempt to include UQ was performed in Reference [4]. It considered the UQ probabilities for the C, P and S outcomes of the three corresponding groups derived from the DL labelling of the saccades and appointed a decision tree (DT) from the mean and standard deviation probabilities calculated for the groups.

This study builds on these previous attempts and now additionally includes the UQ variance information from the MCD-DL model. Since the number of features is now double and exceeds the number of observations, a RF is employed for mining the data set constructed from the mean and standard deviations of the MCD probabilities and variances. Apart from classification, the RF outputs the importance of each statistical variable, as an insight into the process.

The collection of saccadic samples was provided by the Center for Research and Rehabilitacion of Hereditary Ataxias (CIRAH), Holguín, Cuba. It can be accessed for research purposes from the following repository: https://dx.doi.org/10.6084/m9.figshare.11926812.

3. Methods

The current framework for encompassing UQ into the DL architecture through the use of MCD is outlined with respect to the state of the art in the use of this special procedure for this type of medical data. An example of the obtained uncertainty estimates for a chosen register is also shown.

3.1. State of the Art for Mcd-Uq in Biomedical Time Series Modelling

Despite the success of DL for medical problem investigation mostly for images [7–12] but also for biomedical signals [13–16], the question becomes eventually about how trustworthy is the decision support of these networks.

From its introduction by Gal et al. [2] as a Bayesian approximation of a Gaussian process, the use of dropout in DL went beyond controlling overfitting, and directed towards gathering a measure of the certainty of the model on the predicted outcome. Hence, the literature in DL for medicine also shifted to experimenting in this direction as to provide an answer to the earlier question about the degree of trust in the network decision.

Although there is a large number of papers in this UQ direction in the application of DL for medical image processing, the area related to biomedical time series data is not so dense in MCD entries. Besides the higher popularity of image analysis, the MCD results are also more spectacular in visual applications, especially for segmentation. For example, MCD was successfully applied for brain tumor cavity segmentation in Reference [17] and the UQ information proved to be important in the validation of the delineated sections using various DL architectures. MCD was also applied for deep active learning on histological slides, also for segmentation [18]. Another application of MCD for the segmentation of cardiovascular disease images is proposed in Reference [19].

In what follows, some of those few entries that regard MCD-DL specifically for medical time series data analysis are enumerated. Time series in medicine generally define signal data that was collected from sensors: electrocardiography (ECG), electromyography, electrooculography, phonocardiography, blood pressure monitors. Reference [20] proposed a straight CNN and a recurrent version with MCD for pulse detection from ECG signals. The conclusion was that giving feedback to the rescuer only when the uncertainty was below an acceptable threshold significantly increased the balanced accuracy. Reference [21] classifies electromyography hand gesture signals through transfer learning and uses MCD, among batch normalization and early stopping, solely for dealing with overfitting. In Reference [22], a LSTM with MCD was applied to four medical data sets with signal input: ECG for arrythmia classification; phonocardiography for distinguishing between normal and abnormal heart beat; ECG, blood pressure and oxygen saturation in neonatal care; and intracranial and arterial blood pressures, respiratory and heart rates for traumatic injury analysis. The paper demonstrated the benefits of using MCD in increasing classification accuracy in tasks involving medical signals and providing a measure of confidence in the decisions given by the model. On a different note, dropout has been used on recurrent neural networks to improve the robustness of biomedical time series prediction [23].

In this scarcity of studies related to the use of MCD as a UQ approach for investigating biomedical time series data with DL, the present paper advances both a new signal interpretation application from medicine but also takes the MCD technique to measure uncertainty at the higher level of register analysis.

3.2. The Mcd-Dl-Rf Approach for Uq in Register Labeling

The data set with registers of saccades is divided into training, validation and test subsets. The proposed methodology starts from the training saccades that are modelled by a DL (CNN-LSTM) architecture encompassing MCD. The principal goal of LSTM is to keep a long-short term memory in order to handle temporal data. On the other hand, 1D CNN also demonstrated to analyze time series single-handedly [24] or in combination with LSTM [6,25]. Therefore, the architecture used in this study will be a sequence of the two deep networks, with 1D CNN for extracting the features and LSTM for modeling the time dependencies.

The built model is applied to the validation saccades and the uncertainty estimates of class probabilities and variances are used for computing means and standard deviations for each predicted group at the containing register level. The correspondence between the statistical values for the validation registers and the known labels is next learnt by a RF. The reached model classifies the test

registers and additionally outputs a ranking of the importance of the referred statistical variables. The flow is depicted in the diagram of Figure 2.

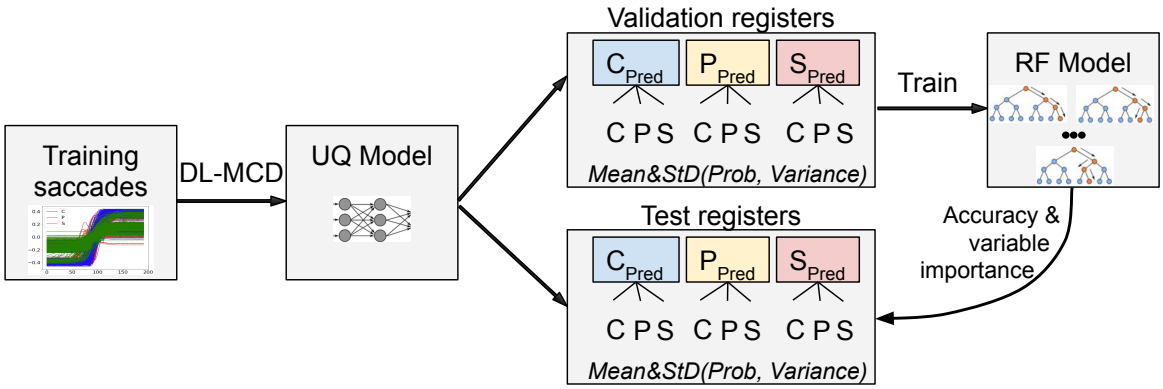

Figure 2. Overview of the proposed methodology.

The approach from training to test through validation is depicted by Algorithm 1. A DL architecture with a MCD layer is trained on the corresponding set of saccades (line 1). The real label for each saccade is given by the training register to which it belongs. Once the DL model is obtained, a sample of N probability outputs for every class (from the N MCD runs) is reached for each saccade from the validation and test registers (lines 3–5). The uncertainty estimates, for example, the average probability and variance, are further computed for every class that the saccade may belong to (lines 6–9). The label with the highest average probability is assigned as the predicted class for the saccade (line 10). This is where the classical MCD procedure ends.

The concept is taken further to the register level, where a different learning approach now continues on the validation instances. A new feature set will be collected from UQ measures computed on the validation registers and will be given to a RF to learn the correspondence between register composition and the real label. First, the number of saccades is recorded for every register (line 13). Then, the saccades within each register are grouped into the predicted labels according to the average probabilities given by the MCD output (lines 14–15). The cardinal of each such found subset is also recorded in the register information (line 16). Subsequently, for each detected group, we compute the mean and standard deviation for the probability and variance of each of the three labels (lines 17–24), resulting in 36 new features for every register. Each validation register is now described by 40 new features that statistically measure the diversity of its interior saccades. A RF is trained on this new register set (line 27). The model is applied on the test registers, described by the same novel features, and the prediction accuracy is calculated (line 28). At the same time, the RF outputs the ranking of these new statistical measurements in connection to the labeling process (line 29).

Algorithm 1: The MCD-DL-RF model: MCD sample output for saccades from DL, computation of the uncertainty estimates over registers and RF construction from the means and standard deviations for register probabilities and variances.

Data: Data set with registers of saccades
Result: RF accuracy and variable importance
1 Learn DL model with a MCD layer on the training saccades - labels given by holding register;
2 **for** *each validation/test saccade* $j = 1, 2, ..., m$ **do**
3 **for** *each MCD run* $i = 1, ..., N$ **do**
4 Obtain class probability sample $Pr_{C_{ij}}$, $Pr_{P_{ij}}$ and $Pr_{S_{ij}}$;
5 **end**
6 **for** $X \in \{C, P, S\}$ **do**
7 Compute average probability $Pr_{X_j} = \frac{1}{N} \sum_{i=1}^{N} Pr_{X_{ij}}$;
8 Compute variance $Var_{X_j} = \frac{1}{N} \sum_{i=1}^{N} (Pr_{X_{ij}} - Pr_{X_j})^2$;
9 **end**
10 Predict label $Pred_j$ = X : $Pr_{X_j} = \max(Pr_{C_j}, Pr_{P_j}, Pr_{S_j})$;
11 **end**
12 **for** *each validation/test register* $i = 1, ..., n$ **do**
13 Calculate the number of saccades $|reg_i|$ (1 feature);
14 Group saccades of register i according to prediction into C^i_{pred}, P^i_{pred}, S^i_{pred}, where
15 X^i_{pred} = {saccade $j \in$ register i : $Pred_j$=X };
16 Compute number of saccades in each subset $|C^i_{pred}|$, $|P^i_{pred}|$, $|S^i_{pred}|$ (3 features);
17 **for** $X \in \{C, P, S\}$ **do**
18 **for** $Y \in \{C, P, S\}$ **do**
19 Compute mean group probability $MeanPr^i_{XY} = \frac{\sum_{j=1}^{|X^i_{pred}|} Pr_{Y_j}}{|X^i_{pred}|}$;
20 Compute StD group probability $StdPr^i_{XY} = \sqrt{\frac{\sum_{j=1}^{|X^i_{pred}|} (Pr_{Y_j} - MeanPr^i_{XY})^2}{|X^i_{pred}|}}$;
21 Compute mean group StD $MeanVar^i_{XY} = \frac{\sum_{j=1}^{|X^i_{pred}|} Var_{Y_j}}{|X^i_{pred}|}$;
22 Compute StD group StD $StdVar^i_{XY} = \sqrt{\frac{\sum_{j=1}^{|X^i_{pred}|} (Var_{Y_j} - MeanVar^i_{XY})^2}{|X^i_{pred}|}}$;
23 **end**
24 **end**
25 Collect resulting: $3(Xs) \times 3(Ys) \times 4(measures) = 36$ features;
26 **end**
27 Build RF from the computed new 40 features of the validation registers;
28 Apply reached RF model on test registers with the same features;
29 Return test classification accuracy and variable importance;

3.3. Visualization of the Output from Mcd-Dl Per Register

This section shows an example of the application of Algorithm 1 for a selected presymptomatic register. Figure 3 shows the obtained average probability and variance results from the MCD runs per possible label for the contained saccades (lines 6–10, 14–15) and Table 1 gives the 36 statistical features computed for the register referring mean and standard deviation from the previous probabilities and variances (lines 17–24).

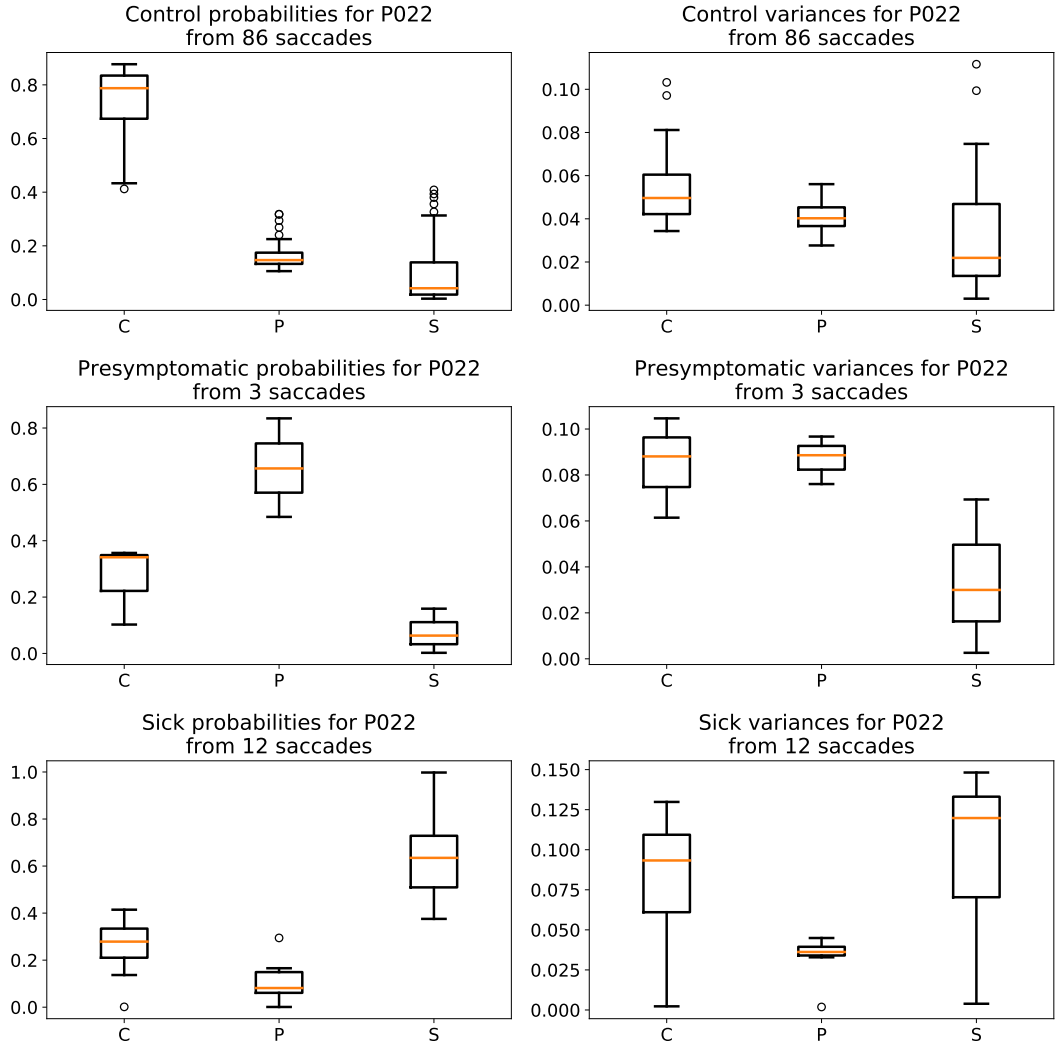

Figure 3. Box plots with probabilities and variances for classes C, P and S reached by the Monte Carlo dropout (MCD) runs for the presymptomatic register P022. Each row corresponds to the saccades that were classified as control (**top**), presymptomatic (**middle**) and sick (**bottom**).

All values in Table 1 are recorded for the register and, together with the amount of saccades labelled with each class in turn, will be further fed to the classifier.

Table 1. Mean and standard deviation values computed from the probabilities and variances for the presymptomatic register P022 that are illustrated in Figure 3.

P022	Probability						Variance					
	Mean			St. Dev			Mean			St. Dev		
	C	P	S	C	P	S	C	P	S	C	P	S
Control	0.75	0.16	0.09	0.12	0.04	0.10	0.05	0.04	0.03	0.01	0.01	0.02
Pres.	0.27	0.66	0.07	0.12	0.14	0.06	0.08	0.09	0.03	0.02	0.01	0.03
Sick	0.26	0.1	0.64	0.11	0.07	0.17	0.08	0.03	0.1	0.04	0.01	0.04

The information presented in the current subsection illustrates that the mere classification of the saccades and the count of how many records from a register are classified into different classes is not enough to determine the label of a register. This justifies the creation of a meta-vector that includes statistics from the values generated by the MCD runs.

4. A Better Comprehension of the Data Via Uq

As each register contains around 70 saccades in average and there are 100 MCD simulations for each such sample in turn, it is natural that for some saccades the variance is higher, while for others it has lower values. Naturally, a higher variance corresponds to a lower degree of certainty. Subsequently, we illustrate in Figure 4 some selected saccades from registers belonging to each of the three classes. The plots from each row show the probabilities and variance results for classes control (first three rows), presymptomatic (rows 4–5) and sick (last row).

The first column of the plots shows on the horizontal the probability that the saccade is classified as control, the middle column corresponds to the prediction for class presymptomatic and the right one to a label of sick. When the bars have a large spread, there is a high variance in the results of the 100 MCD runs, while when they are gathered closer, the variance is low. The vertical axes show the number of MCD runs in which the probabilities on the horizontal axis where the bars lie are achieved.

For the control class the results for three saccades are shown in the plots from the first three rows. The first saccade is classified as control with a probability of 0.78 and a relatively low variance. It has however a probability of 0.19 of being labeled as presymptomatic. The next two saccades from the control register for which the results are illustrated below have an increased variance and the results seem more ambiguous, although the largest probability falls into the control class in both situations. There are also saccades in control registers that are more clearly classified as control, with a higher probability for control and a lower variance, but such examples were not included in the article due to space economy.

The results for two distinct saccades in the presymptomatic register P022 are illustrated in the subsequent two rows in Figure 4. While the first one is assigned to the presymptomatic class with a high probability of 0.83, the next one has the predicted label balanced between sick (probability 0.56) and control (0.35) with a very high variance in both cases. Nevertheless, it has to be stated that presymptomatic registers contain many saccades where the probability for control dominates in the MCD simulation results.

Lastly, we have an example with a saccade that is clearly assigned to the sick class (last row in Figure 4). Again, there are also other saccades where the distinction is not so clear, probabilities for the other classes are higher and the variance has larger values.

Figure 5 illustrates an overview for the probabilities obtained when applying the 100 MCD simulations for all saccades from the three registers chosen as representatives for the distinct classes. Each bar in the plots represents the probability for a saccade to be assigned to either control (blue), presymptomatic (orange) or sick (red). The two plots on each row correspond to the cases when there was a variance for all classes below a certain threshold of 0.04 on the left side or above 0.03 on the right. The thresholds are chosen such that there exists a balanced distribution of bars in all cases. Note that the same threshold values are kept for all three registers (and classes, respectively), but the balance between the saccades with lower variance (left plots) and the ones with higher variance (right plots) is very different. For instance, for almost 79% of the sick registers the classification of the MCD runs led to low variance (i.e., cases like the last row in Figure 4), while for the control and presymptomatic cases there are less than 13% where the model had such a low variance, and high certitude, respectively. This also offers an explanation why it is harder to distinguish between the control and presymptomatic classes, while sick is easier to delineate [5,6].

The saccades for which results are illustrated in Figure 4 belong to the registers that are shown in Figure 1 and the findings for some of their saccades are also shown in Figure 5—for instance, saccade 1134 for which the probabilities per class and variances are represented in the second row in Figure 4 also appears as the first group of three bars in the second plot, first row of Figure 5.

It is interesting to observe that the patterns concerning the probabilities for the distinct classes change when the variance is higher or lower. This can be seen especially for the first and third rows in the figure.

Although we state that the three registers for which various results are illustrated in Figures 4 and 5 are representative for the three classes of the problem, we underline that the same clear pattern is not observed in all registers.

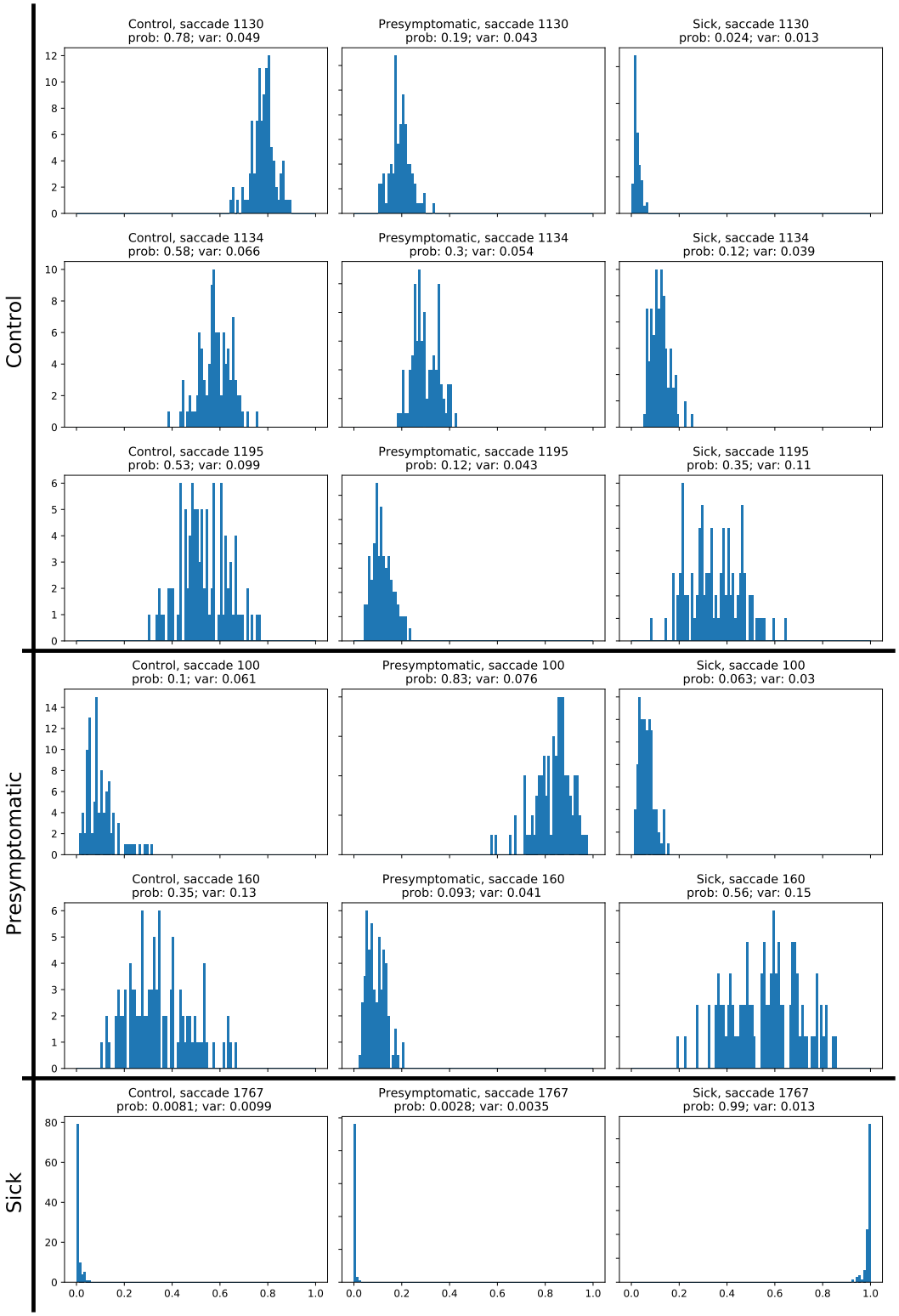

Figure 4. Example of the classification probabilities and variances from the 100 MCD runs for some saccades from class control (3 saccades in register C029, rows 1–3), from class presymptomatic (2 samples from register P022, rows 4–5) and from class sick (1 saccade from register S021, last row).

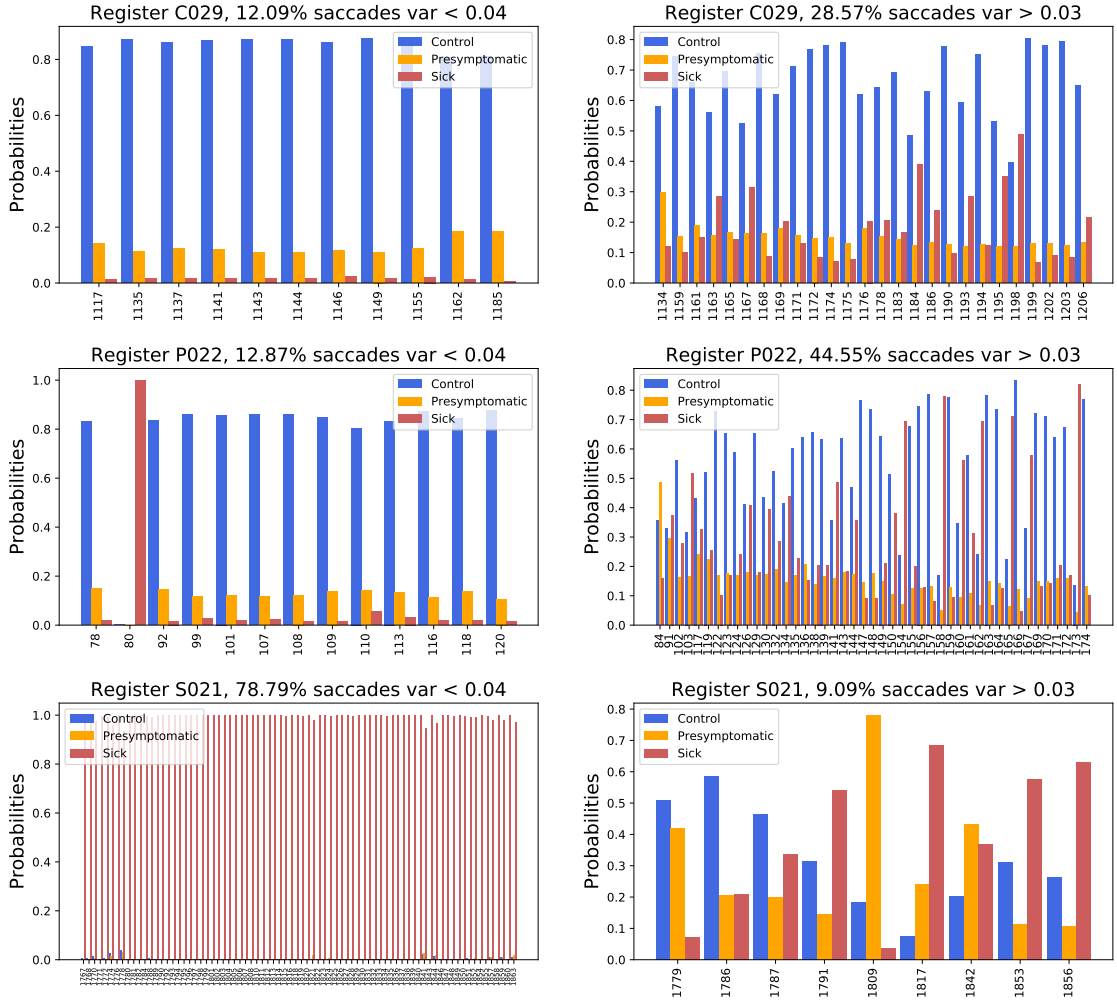

Figure 5. The label prediction with probabilities and variances for each saccade of the example registers as accumulated from the 100 MCD runs.

5. Experimental Results

The current section presents the setup used for the application of the proposed methodology, the obtained classification results along with the most important features as discovered by the RF and further discussions. The next subsection presents more attempts tried during pre-experimentation that did not eventually lead to the most successful results.

5.1. Pre-Experimental Observations

In one of the tried scenarios, we tested the idea that there might exist only two types of saccades, specific for control and for sick, respectively, and the register label might be established by the balance of the amount of saccades having each of these two labels. Accordingly, we tried an approach where only saccades from registers that belong to these 2 classes, control and sick, are used for training the CNN-LSTM. The model would then classify the validation and test saccades into these two classes. The shallow classifier would subsequently use the outputs computed by the MCD probabilities and variances for these two classes to learn to predict the class of the registers. The idea of putting such a scenario into practice came from the relatively lower importance that was found for the features derived from the presymptomatic determined saccades. The importance is discussed in detail in Sections 5.3 and 5.4. However, the classification results were down by 2% as opposed to the case where all three classes are considered at the training level of the saccades and therefore the case with all three labels was used throughout the experiments.

We tried several shallow classifiers for distinguishing the classes of the registers based on the statistical UQ features recorded for them. One option was support vector machines (SVM), both linear and with a radial kernel. The data was normalized before applying SVM, but in both settings all the presymptomatic registers were mistaken for either control or sick classes. Besides SVM, decision trees were also tried, as they have shown good results previously [4]. They led to better outputs than SVM and these are further outlined in the next subsection.

Another attempt addressed the inclusion of a feature selection method represented by a hill climbing algorithm to pick the attributes derived from the MCD simulations for the registers (the 40 features). The evaluation of the hill climbing was given by the classification accuracy reported by the RF classifier on a validation set. Accordingly, the features that led to the best classification accuracies on the validation set were used for the test set as well. However, the final results were very similar to the ones obtained by applying the RF directly on the data set containing all features. This goes in line with the observed trait of RF to be able to handle multidimensional data well.

5.2. Experimental Setup

The registers are split into training, validation and test sets such that the distribution of classes are maintained balanced in each such set. The same split was used in previous works [4,6] and is kept here as well to make a fair comparison. The DL architecture consists of 2 CNN layers of size 3 and 128 filters and one LSTM layer of 100 units. The MCD steps in with a dropout rate of 0.5. The implementation is done in Python using the TensorFlow library and running on the GPU.

The number of Monte-Carlo applications was varied during pre-experimentation from 30 up to 500 but we noticed that the improvement for the classification accuracy for the saccades of the ensemble model stopped when the number of MCD applications went beyond 100 (actually on 83, but we selected a more round number). In a previous study [4], we kept this number at 500, which was consuming computations, without any gain in the accuracy of the ensemble. There are 10 distinct runs of the MCD approach on the saccades data set and they subsequently create 10 separate numerical data sets on which different RF instances will be trained.

RF is applied 10 times on each data set and the mean results are reported. The results are further accumulated from the 10 data sets and the mean from these is reported as the final result. The RF implementation uses the Scikit-learn package in Python [26]. As concerns the parameter settings of the algorithm, the default 100 trees are used, each with a maximum depth of 3. The quality of the split uses the default Gini impurity criterion.

The values for the involved parameters are generally kept to simple and default settings. Nevertheless, some metaheuristics could be used to make fine tuning especially for the deep learning architecture [27].

5.3. Results and Visualization

Table 2 shows the results obtained by the RF or DT using either the data with the entire set of features that include those related to variance as well and only the UQ probabilities. Beside the classification accuracy, the F1 score and the running time are computed. The statistical tests on the last two rows point the differences between the most prolific approach and the others used in comparison.

Table 2. Classification results, running time and statistical tests for the proposed approach against other setups, including decision tree (DT) without variance [4] in the last column.

Measure	**RF with Var**	**RF no Var**	**DT with Var**	**DT no Var**
Accuracy	88.59	86.88	83.12	82.88
F1 score	87.75	85.52	83	82.77
Running time (sec)	0.38	0.39	0.006	0.006
Statistical tests, RF with Var vs. others				
Wilcoxon signed-rank test	-	6.05×10^{-4}	5.9×10^{-11}	2.6×10^{-8}
Paired Student's t-test	-	2.8×10^{-3}	1.47×10^{-12}	2.79×10^{-10}

Figure 6 illustrates the confusion matrices for the same 4 compared approaches in Table 2.

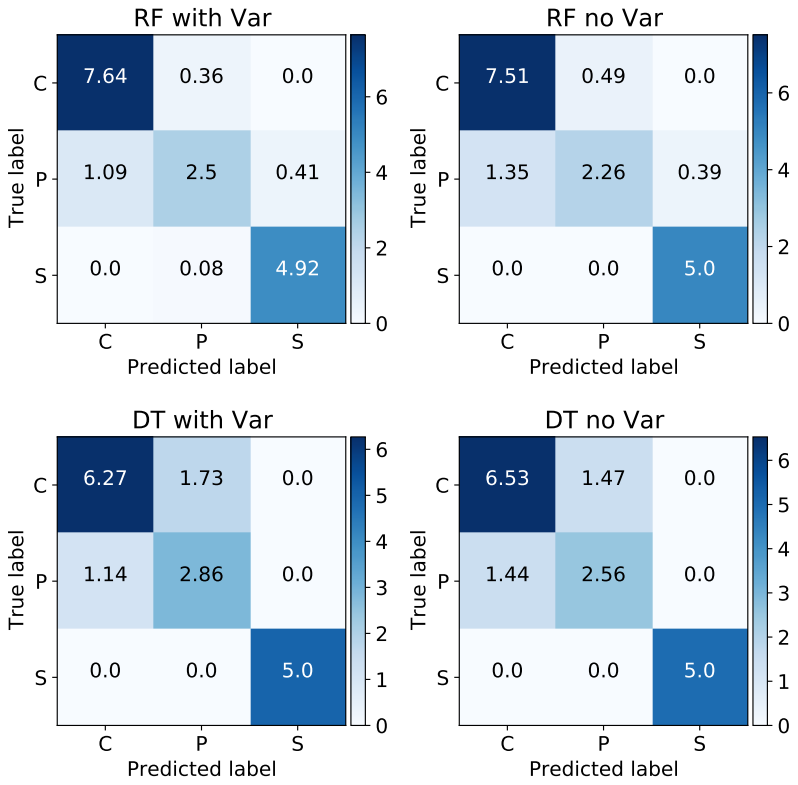

Figure 6. Confusion matrices for Random Forest (RF) with and without variances (on the top row) and the corresponding application of DT (on the bottom row).

Figure 7 shows the importance of each feature based on the Gini impurity that is used in the nodes of the trees in the RF.

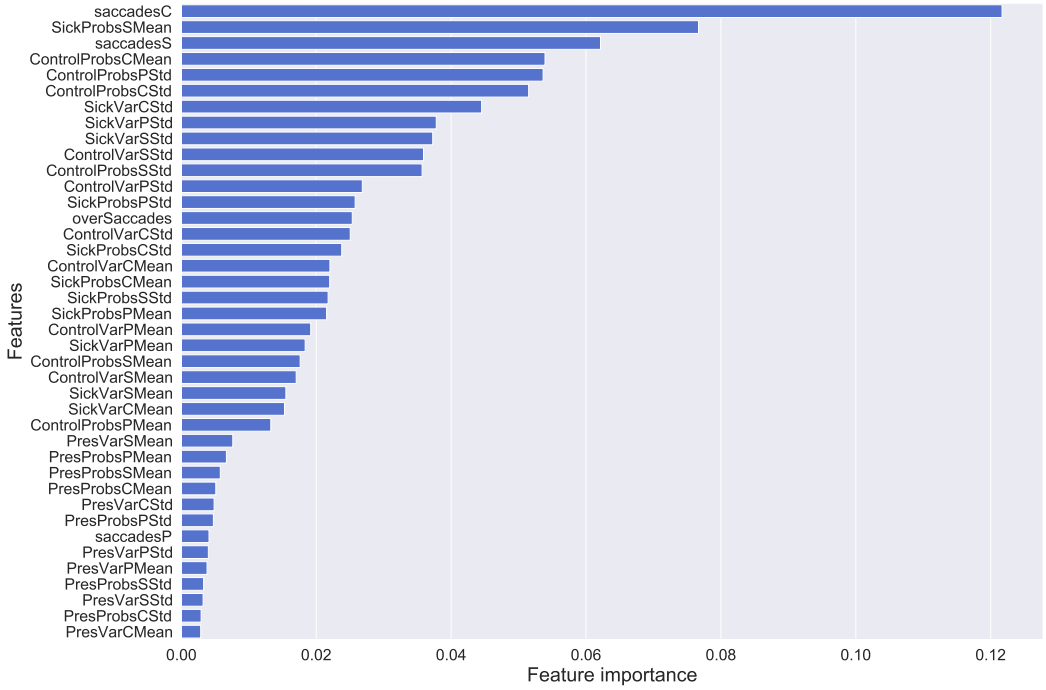

Figure 7. Importance of the features as derived from Gini impurity.

Figure 8 illustrates the importance of the features along with the standard deviation for each class in turn, following the method proposed by Saabas in Reference [28].

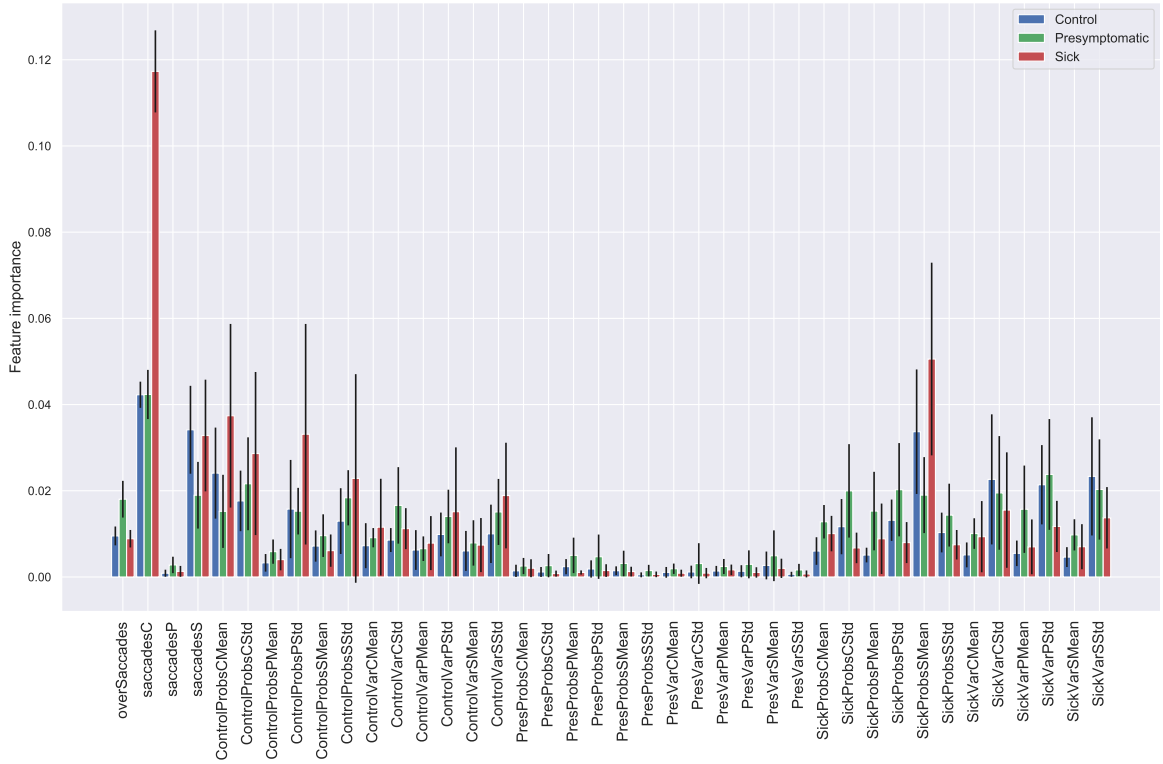

Figure 8. Importance of the features for each class in turn.

5.4. Discussion

The statistical tests were computed by comparing each individual run from the 10 repeated applications of the classifier on every one of the 10 distinct data sets obtained from the MCD applications. Consequently there are 100 distinct classification accuracies for each of the 4 classifiers in Table 2 used for calculating the *p*-values in the last 2 rows of the table.

RF clearly led to better results as compared to DT, as the results were better regardless of using the features related to variance or not. While the variance attributes did not add value for the DT model, that is, the classification accuracy of 83.12% with variance versus 82.88% without it, the importance of the same attributes proved significant for the RF classifier. This answers the research question that was formulated in the introduction: the incorporation of the variance was not enough to improve the results from Reference [4], the DT classifier also had to be changed to a RF to take advantage of the higher amount of information. Both Wilcoxon signed-rank test and the paired Student *t*-test obtained *p*-values that were below 0.05, rejecting thus the null hypothesis, and proving thus that the samples belong to populations with different distributions and hence the improvement in classification accuracy is significant.

Figure 6 illustrates how the test registers are labeled in each of the 4 compared situations. It is worth underlying that in the DT classification the results of the sick class did not interfere in any way with the other 2: neither the control nor the presymptomatic registers are mistaken for sick or vice-versa. The RF mistakes some of the presymptomatic registers with the sick in both cases, when the variance is used or let out. There is even a very small amount of cases when some sick registers are mistaken for presymptomatic when RF with variance is employed. Nevertheless, the latter performs a more accurate prediction for the control registers.

Figures 7 and 8 illustrate the importance of the involved features as they are derived from the Gini impurity in the former case and as specialized on the registers and classes in the latter. Since impurity-based feature importance favors features with many unique values, we additionally used an approach that shows the importance for each individual sample in turn. Saabas proposed the method to interpret the predictions of the RF for each record [28]. In order to provide a unified plot, we gathered all registers that belong to the same class and computed the averages over multiple runs. The most important feature is indicated the same in both figures, that is the number of control saccades in the register. The plot in Figure 8 additionally illustrates that this attribute helps especially in assigning the register class to sick, since this is the largest bar in the chart. However, the attribute remains very important for the other 2 classes, as well. The second most important feature in Figure 7, the $MeanPr_{SS}$ again remains in agreement with the importance detected by the model where the discrimination is made between classes: this is the second most important attribute in the entire plot and is again crucial for the sick class (red bar). The two plots also agree that the attributes related to the presymptomatic class have less importance. They are the ones having the name starting with *Pres* and are situated at the bottom of the plot in Figure 7 and are at the small bars in the middle of the chart in Figure 8. Nevertheless, when we tried to remove the presymptomatic class from the training saccades, and subsequently from the registers in the second part of the methodology, the results were weaker than the currently reported ones, as also described in Section 5.1.

6. Conclusions

The methodology targeted in the current study started from the intricacy of the medical task concerning the classification of registers containing saccades from patients into three classes: control (or healthy), presymptomatic and sick. The complexity derives from the fact that not all saccades have a similar nature: the control registers contain some that have an unusual (distinct from healthy) behaviour, a shape that usually occurs in the presymptomatic registers. There seems to be a thin delineation between the two classes and the solution put forward in this research makes use of MCD for UQ to better discriminate between them. The registers from the third class, sick, are easier to distinguish, since a majority of saccades have a non-healthy appearance.

It is not the first time UQ is applied for the current task, a previous study led to the results in Reference [4]. However the current research does a more extended work, which encompasses a large variety of tried settings. The important steps forward come from the following observations: (1) a smaller number of MCD steps is enough to reach similar results; (2) the variance from the MCD steps brings additional useful information and (3) RF proves to be a more prolific classifier for the new data set created from the calculated statistical features from the MCD uncertainty estimates. This also answers the research question addressed in the introduction: there were two changes necessary (use of variance and RF instead of DT) as opposed to Reference [4] in order to make a important step as concerns the accuracy. Moreover, other approaches that led to competitive results are studied and summarized, like employing a wrapper feature selection prior to using the RF or even removing completely the presymptomatic class from saccade classification, but they still proved to be significantly weaker than the ones obtained by the proposed tandem. Other statistics besides mean and standard deviation might also arise and contain useful information from the probabilities and variances, but these would further increase the number of features and would probably necessitate some means of reducing the number of attributes like using some correlation analysis as in Reference [29] or even some mechanism inside RF as in Reference [30]. On the other hand, the integration of other clinical information about the patients besides the saccades themselves [31] might additionally increase the quality of the diagnosis. An important parameter to be studied is the MCD rate: its value in the current work was empirically established to 0.5, but its influence over the final results might be crucial and its fine tuning might lead to better accuracy.

From the theoretical viewpoint, the paper generalizes the MCD approach for UQ by bringing it to a higher grouping level of records (i.e., registers of trials) and demonstrates its usefulness for DL support in medicine.

Author Contributions: Conceptualization, R.S. and C.S.; methodology, C.S. and R.S.; software, C.S.; validation, M.A. and G.J.; formal analysis, R.S.; resources, R.R.-L.; writing—original draft preparation, R.S. and C.S.; writing—review and editing, M.A. and G.J.; visualization, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Spanish Ministry of Science and Innovation, through the Plan Estatal de Investigación Científica y Técnica y de Innovación, Project TIN2017-88728-C2-1-R, and the University of Málaga-Andalucía-Tech through the Plan Propio de Investigación y Transferencia, Project DIATAX: Integración de nuevas tecnologías para el diagnóstico temprano de las Ataxias Hereditarias.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
UQ	Uncertainty Quantification
MCD	Monte Carlo Dropout
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
RF	Random Forest
DT	Decision Tree
HC	Hill Climbing

References

1. Bacciu, D.; Lisboa, P.J.; Martin, J.D.; Stoean, R.; Vellido, A. Bioinformatics and medicine in the era of deep learning. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'18), Bruges, Belgium, 25–27 April 2018; Verleysen, M., Ed.; pp. 345–354.
2. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1050–1059.
3. Vishnu, T.; Diksha, G.; Malhotra, P.; Vig, L.; Shroff, G. Data-driven Prognostics with Predictive Uncertainty Estimation using Ensemble of Deep Ordinal Regression Models. *Int. J. Progn. Health Manag.* **2019**, *10*, 027.
4. Stoean, R.; Stoean, C.; Abdar, M.; Atencia, M.; Velázquez-Pérez, L.; Khosravi, A.; Nahavandi, S.; Acharya, U.R.; Joya, G. Automated Detection of Presymptomatic Conditions in Spinocerebellar Ataxia Type 2 using Monte-Carlo Dropout and Deep Neural Network Techniques with Electrooculogram Signals. *Sensors* **2020**, *20*, 3032. [CrossRef]
5. Stoean, C.; Stoean, R.; Becerra-García, R.A.; García-Bermúdez, R.; Atencia, M.; García-Lagos, F.; Velázquez-Pérez, L.; Joya, G. Unsupervised Learning as a Complement to Convolutional Neural Network Classification in the Analysis of Saccadic Eye Movement in Spino-Cerebellar Ataxia Type 2. In *Advances in Computational Intelligence*; Rojas, I., Joya, G., Catala, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 26–37.
6. Stoean, R.; Stoean, C.; Becerra-García, R.A.; García-Bermúdez, R.; Atencia, M.; García-Lagos, F.; Velázquez-Pérez, L.; Joya, G. A Hybrid Unsupervised - Deep Learning Tandem for Electrooculography Time Series Analysis. *PLoS ONE* **2020**, submitted.
7. Abdar, M.; Makarenkov, V. CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement* **2019**, *146*, 557–570. [CrossRef]
8. Stoean, R. Analysis on the potential of an EA–surrogate modelling tandem for deep learning parametrization: an example for cancer classification from medical images. *Neural Comput. Appl.* **2020**, *32*, 313–322. [CrossRef]
9. Mittal, S.; Stoean, C.; Kajdacsy-Balla, A.; Bhargava, R. Digital Assessment of Stained Breast Tissue Images for Comprehensive Tumor and Microenvironment Analysis. *Front. Bioeng. Biotechnol.* **2019**, *7*, 246. [CrossRef]
10. Benhammou, Y.; Achchab, B.; Herrera, F.; Tabik, S. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing* **2019**, *375*, 9–24. [CrossRef]
11. Sahlsten, J.; Jaskari, J.; Kivinen, J.; Turunen, L.; Jaanio, E.; Hietala, K.; Kaski, K. Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading. *Sci. Rep.* **2019**, *9*. [CrossRef]
12. Yang, X.; Tang, W.T.; Tjio, G.; Yeo, S.Y.; Su, Y. Automatic detection of anatomical landmarks in brain MR scanning using multi-task deep neural networks. *Neurocomputing* **2019**, *396*, 514–521. [CrossRef]
13. Plawiak, P.; Acharya, U.R. Novel deep genetic ensemble of classifiers for arrhythmia detection using ECG signals. *Neural Comput. Appl.* **2019**, 1–25. [CrossRef]
14. Yildirim, O.; Baloglu, U.B.; Tan, R.S.; Ciaccio, E.J.; Acharya, U.R. A new approach for arrhythmia classification using deep coded features and LSTM networks. *Comput. Methods Programs Biomed.* **2019**, *176*, 121–133. [CrossRef] [PubMed]
15. Alfaras, M.; Soriano, M.C.; Ortín, S. A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. *Front. Phys.* **2019**, *7*, 103. [CrossRef]
16. Ledezma, C.A.; Zhou, X.; Rodríguez, B.; Tan, P.J.; Díaz-Zuccarini, V. A modeling and machine learning approach to ECG feature engineering for the detection of ischemia using pseudo-ECG. *PLoS ONE* **2019**, *14*, e0220294. [CrossRef]
17. Jungo, A.; Meier, R.; Ermis, E.; Herrmann, E.; Reyes, M. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. *arXiv* **2018**, arXiv:1806.03106.
18. Lubrano di Scandalea, M.; Perone, C.S.; Boudreau, M.; Cohen-Adad, J. Deep active learning for axon-myelin segmentation on histology data. *arXiv* **2019**, arXiv:1907.05143.
19. Guo, F.; Ng, M.; Goubran, M.; Petersen, S.E.; Piechnik, S.K.; Neubauer, S.; Wright, G. Improving Cardiac MRI Convolutional Neural Network Segmentation on Small Training Datasets and Dataset Shift: A Continuous Kernel Cut Approach. *Med. Image Anal.* **2020**, *61*, 101636. [CrossRef]

20. Elola, A.; Aramendi, E.; Irusta, U.; Picón, A.; Alonso, E.; Owens, P.; Idris, A. Deep Neural Networks for ECG-Based Pulse Detection during Out-of-Hospital Cardiac Arrest. *Entropy* **2019**, *21*, 305. [CrossRef]
21. Côté-Allard, U.; Fall, C.L.; Drouin, A.; Campeau-Lecours, A.; Gosselin, C.; Glette, K.; Laviolette, F.; Gosselin, B. Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 760–771. [CrossRef]
22. van der Westhuizen, J.; Lasenby, J. Bayesian LSTMs in medicine. *arXiv* **2017**, arXiv:1706.01242.
23. Bacciu, D.; Crecchi, F. Augmenting Recurrent Neural Networks Resilience by Dropout. *IEEE Trans. Neural Networks Learn. Syst.* **2020**, *31*, 345–351. [CrossRef] [PubMed]
24. Stoean, C.; Paja, W.; Stoean, R.; Sandita, A. Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations. *PLOS ONE* **2019**, *14*, e0223593. [CrossRef] [PubMed]
25. Zhu, F.; Ye, F.; Fu, Y.; Liu, Q.; Shen, B. Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. *Sci. Rep.* **2019**, *9*. [CrossRef] [PubMed]
26. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
27. Bacanin, N.; Bezdan, T.; Tuba, E.; Strumberger, I.; Tuba, M. Optimizing Convolutional Neural Network Hyperparameters by Enhanced Swarm Intelligence Metaheuristics. *Algorithms* **2020**, *13*, 67. [CrossRef]
28. Saabas, A. Treeinterpreter, 2018. Python Package, Version 0.2.2. Available online: http://blog.datadive.net/interpreting-random-forests/ (accessed on 1 June 2020).
29. Kochev, S.; Stevchev, N.; Kocheva, S.; Eftimov, T.; Simjanoska, M. A Novel Approach for Modelling the Relationship between Blood Pressure and ECG by using Time-series Feature Extraction. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020), Valletta, Malta, 24–26 February 2020; Vilda, P.G., Fred, A.L.N., Gamboa, H., Eds.; Volume 4, pp. 228–235. [CrossRef]
30. Paja, W.; Pancerz, K.; Grochowalski, P. Generational Feature Elimination and Some Other Ranking Feature Selection Methods. In *Advances in Feature Selection for Data and Pattern Recognition*; Springer International Publishing: Cham, Switzerland, 2018; pp. 97–112. [CrossRef]
31. Morales, J.C.; Carrillo-Perez, F.; Castillo-Secilla, D.; Rojas, I.; Herrera, L.J. Enhancing Breast Cancer Classification via Information and Multi-model Integration. In *Bioinformatics and Biomedical Engineering*; Rojas, I., Valenzuela, O., Rojas, F., Herrera, L.J., Ortuño, F., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 750–760.