Dear reviewers,

We highly appreciate all your suggestions and comments. We have used your feedback to improve the form and content of our document. The responses to your questions are registered below your statements. The actions taken are recorded in this document and performed in the manuscript; changes and additions are marked in bold font.

Again, thanks for your time.

Reviewer 1.

Table 2 shows the performance of the neuronal network model. However, an additional table is also needed that should compare the neuronal network model presented in this paper with other current state-of-the-art models. I recommend showing the accuracy, recall, F1-score, and precision of this neuronal network model and the other compared previous models.

R/ Thanks for these suggestions. The accuracy, recall, and F1 score are metrics for classification solutions. They all use the terms of the confusion matrix, and their domain is the discrete world. In the context of this work, translating the pulses to a volume is better framed by regressors since the output must be a number in the continuous domain. In this case, the Mean Absolute Error (MAE) is preferred. The Mean Squared Error (MSE) is also used, but physicians used to have a better sense of accuracy with the MAE since it does not involve quadratic operators.

We followed your suggestion and tested other state-of-the-art AI approaches. They are:

5.1 Linear regression

```
# getting all columns except lastone which is the supervising values
features = df2.iloc[:,:-1]
# put everything in an array
X = features.values
# geeting the supervisor values
supvisor = df2.iloc[:,-1]
# put the supervisor values in an array
y = supvisor.values
# preparing validation
Xtrain_lr, Xtest_lr, ytrain_lr, ytest_lr = train_test_split(X, y, test_size = .3, random_state = 0)
# instantiating the linear regression
lr = LinearRegression()
# getting the prediction model
lr.fit(Xtrain_lr, ytrain_lr)
# predicting
ypred_lr = lr.predict(Xtest_lr)
# metrics
mse_lr = mean_squared_error(ytest_lr, ypred_lr)
mae_lr = mean_absolute_error(ytest_lr, ypred_lr)
print('Mean squared error in LR: ', mse_nn)
print('Mean absolute error in LR: ', mae_nn)
```

executed in 3ms, finished 15:31:44 2023-02-07

5.2 Polynomial regression

```
# getting all columns except lastone which is the supervising values
features = df2.iloc[:,:-1]
# put everything in an array
X = features.values
# geeting the supervisor values
supvisor = df2.iloc[:,-1]
# put the supervisor values in an array
y = supvisor.values
#creating the model
poly = PolynomialFeatures(degree=4)
poly_train = poly.fit_transform(X_train)
poly_test = poly.fit_transform(X_test)
regression = linear_model.LinearRegression()
model = regression.fit(poly_train, y_train)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = .3, random_state = 0)
# predicting
y_pred = model.predict(poly_test)
# metrics
mse_pr = mean_squared_error(y_test, y_pred)
mae_pr = mean_absolute_error(y_test, y_pred)
print('Mean squared error in PR: ', mse_pr)
print('Mean absolute error in PR: ', mae_pr)
```

executed in 3ms, finished 15:35:55 2023-02-07

5.3 Neuronal network

```
# shuffle data
df2_shu = df2.sample(frac = 1)
# getting all columns except lastone which is the supervising values
features = df2.iloc[:,:-1]
# put everything in an array
X = features.values
# geeting the supervisor values
supvisor = df2.iloc[:,-1]
# put the supervisor values in an array
y = supvisor.values
# divide the data into train and test datasets
Xtrain_nn, Xtest_nn, ytrain_nn, ytest_nn = train_test_split(X, y, test_size = .30, random_state = 0)
# normalizing
scaler = preprocessing.StandardScaler().fit(X)
Xtrain_nn_scaled = scaler.transform(Xtrain_nn)
Xtest_nn_scaled = scaler.transform(Xtest_nn)
# NN model
model = Sequential()
model.add(Dense(23, input_dim = 23, activation = 'relu'))
model.add(Dense(12, activation = 'relu'))
model.add(Dense(10, activation = 'relu'))
model.add(Dense(1, activation = 'linear'))
# loss, optimizer and metrics
model.compile(loss = 'mean_squared_error', optmizer = 'adam', metrics = ['mae'])
model.summary()
# fitting
history = model.fit(Xtrain_nn_scaled, ytrain_nn, validation_split= 0.3, epochs = 4000)
# fair format metrics
mse_nn, mae_nn = model.evaluate(Xtest_nn_scaled, ytest_nn)
print('Mean squared error: ', mse_nn)
print('Mean absolute error: ', mae_nn)
```

executed in 13ms, finished 15:49:08 2023-02-07

5.4 Decision tree

```
tree = DecisionTreeRegressor()
tree.fit(Xtrain_nn_scaled, ytrain_nn)
ypred = tree.predict(Xtest_nn_scaled)
mse_dt = mean_squared_error(ytest_nn, ypred)
mae_dt = mean_absolute_error(ytest_nn, ypred)
print('Mean squared error with decision tree: ', mse_dt)
print('Mean absolute error with decision tree: ', mae_dt)
```

executed in 5ms, finished 15:49:23 2023-02-07

5.5 Random forest

```
model = RandomForestRegressor(n_estimators = 10, random_state = 30)
model.fit(Xtrain_nn_scaled, ytrain_nn)
y_pred_RF = model.predict(Xtest_nn_scaled)
mse_RF = mean_squared_error(ytest_nn, y_pred_RF)
mae_RF = mean_absolute_error(ytest_nn, y_pred_RF)
print('The mean squared error in RF is {}'.format(mse_RF))
print('The mean absolute error in RF is {}'.format(mae_RF))
```

executed in 18ms, finished 15:51:07 2023-02-07

The MAE of the listed regression methods is presented in the following table. We will include this table in the results section of the paper.

Folding	**Regression models** (23 features, 44 formulations, split = 0.3)				
	Linear regression	**Polynomial regression**	**Neuronal Network**	**Decision tree**	**Random forest**
		polynomial degree: 2	Loss: MSE, Optimizer: Adam, nodes per layer: 23,12,10,1, activation per layer: relu, relu, relu, linear	Type: regressor	Type: regressor, Estimators: 10
	MAE	**MAE**	**MAE**	**MAE**	**MAE**
1	1578.54	3260.24	129.99	1908.61	578.12
2	1876.58	4125.73	79.71	1155.41	411.65
3	1599.33	3518.47	85.81	1196.66	388.96
Average	**1684.81**	**3694.81**	**98.50**	**1420.22**	**459.57**

Although these methods are comparable by MAE, the comparison remains unfair since there is space to configure them to improve their performance independently. It would be interesting to see how the simplistic linear and polynomial regressions would behave when optimized. The presented development will be enhanced if the simplistic approaches reach a good accuracy (low MAE) because it would make the embedded hardware run faster.

The optimization of other methods was not executed due to the excellent performance of the neuronal network. Recall that the smallest volume read with a caliper yielded an analytical volume of 7329 mm3, and the worst obtained MAE is 129.99 mm3 which is 1.77% of the smallest measured volume. Therefore, the WD is certified to measure volumes by water displacement with high precision.

Reviewer 2.

- Dear authors, it is an interesting paper about segmentation from MRI. I have the following suggestions:

Title: please change for: Eliminating the need for manual segmentation to determine the size and volume of lateral ventricles images from MRI

R/ Thanks for your suggestion. We have proposed this method as a general mechanism to determine errors attained by manual segmentation and consequently discourage its use in medicine. Segmenting the ventricles is proof of concept, but we could have performed the same in any other body part. We are interested in keeping it general. We propose to set the title as Eliminating the Need for Manual Segmentation to Determine Size and Volume from MRI. A proof of concept on Segmenting the Brain Lateral Ventricles.

Abstract: at the first Page of the paper please review the numbers: 2\%

R/ 2% of a discrepancy between the AVVE automation and the gold standard was the limit we proposed to validate the operation of the AVVE. Therefore, the number is correct.

The other percentages are the highest discrepancies between the volumes measured by human operators and the certified automation, and they are also correct.

The errors are a percentages of the volume measured with a certified tool.

The conclusions of the abstract is not so clear as the conclusion of the paper. Please, review the conclusion of the abstract.

R/ That is true. We have changed the last paragraph of the abstract, and now it reads as follows.

"

The errors induced are large enough to adversely affect decisions that may lead to less-than-optimal treatments; therefore, we suggest avoiding manual segmentation whenever possible.

Introduction: please explain in few words the reference 14.

R/ We have written a synopsis of the method, which appears in bold fonts in the new version of the article. It reads as follows:

"

This study examined the accuracy of manual segmentation for ventricular volume (3D) and compared it to a certified version of the Automatic Ventricular Volume Estimator (AVVE), a method we developed in [14]. The AVVE uses Support Vector Machine (SVM) to classify the voxels belonging to volumes of interest automatically. This statistical estimator receives four features extracted from the studied image and the ventricular masks as a supervisory factor. When presented to the research community, the AVVE was validated using manually segmented masks, but in this delivery, the AVVE has been certified for accuracy using a reproducible pipeline. Then, with the certified AVVE, we measure and report the errors attained by human operators while segmenting the lateral ventricles.

"

Methods: Please include the abbreviations of the Figures and Tables at the legend after each Figure and Table.

R/ We have expanded all abbreviations used in figures and tables. Please observe the bold fonts in every updated caption.

Results: the interoperator measurements errors are up to 50%. What are the explanations for such number and other high Numbers errors?

R/ Such a considerable error rate is often read in significant volumes where there is more chance to make mistakes due to more extended boundaries. Also, larger structures are more affected by partial volume effects. In general, regardless of the errors' nature (big or small) concerning a gold standard in this artisan activity can be only explained by human factors.

Discussion: please compare the findings of this study with other studies of automation measurements in MRI.

R/ The article aims to quantify and report human errors during segmentation tasks. We selected the lateral ventricles (LV) because our team has significant experience with these structures. The LV creates a good contrast in MRI and CT, even in low-quality acquisitions, facilitating the reproducibility of our methods. We could not find any other paper reporting manual segmentation errors that referred to a reliable gold standard while measuring LV in children or a different structure in any other type of subject.

Papers report LV volumes [Melhem et al. 2000; Sarı et al. 2015] but their methods use manual segmentation or indirect mechanism such as the Evans' index; therefore, there is no shared space for comparison. Some reported volumes in [Melhem et al. 2000] may match the age ranges that we register in this manuscript; however, their patients have a brain malformation different from hydrocephalus, which is the only abnormality we report. Other authors declare VL volumes of various pathologies [Del Re et al. 2016; Ertekin et al. 2016; Turner,Greenspan & van Erp 2016], based on manual segmentation.

The problem of validating automatic and semi-automatic tools with manual assessments in medicine has been underrated. Nevertheless, some authors have recently spoken out about the inconsistency of using unstable manual segmentation as a grand truth and proposed to believe in the machine's capacity to learn and be reproducible [Zhang et al. 2020] for accomplishing tasks with precision. [Zhang et al. 2020] justified their efforts with a 10% discrepancy between operators in a multiple-sclerosis framework while segmenting brain structures. However, reporting the differences between operators obviates the target and, thus, precision. In other words, both operators could be in the same numbers and far away from the real numbers. Losing the target is a natural result when we lack an objective gold standard. This missing part propagates the hesitations to the scenario where the artificial intelligence machine performs the segmentation. Is it obtained the correct numbers? How can we ensure that? Still, we can not compare our findings with anything reported before because we propose the creation of a gold standard, something missing in the 8.880 entries displayed by google scholar after the search string "Segmentation algorithms in medical imaging" only in 2023.

We will include this explanation and references in the article's discussion section.

References

Del Re, E. C.; Konishi, J.; Bouix, S.; Blokland, G. A.; Mesholam-Gately, R. I.; Goldstein, J.; Kubicki, M.; Wojcik, J.; Pasternak, O.; Seidman, L. J. and others (**2016**). *Enlarged lateral ventricles inversely correlate with reduced corpus callosum central volume in first episode schizophrenia: association with functional measures*, Brain imaging and behavior 10 : 1264-1273.

Ertekin, T.; Acer, N.; Köseoğlu, E.; Zararsız, G.; Sönmez, A.; Gümüş, K. and Kurtoğlu, E. (**2016**). *Total intracranial and lateral ventricle volumes measurement in Alzheimer's disease: A methodological study*, Journal of Clinical Neuroscience 34 : 133-139.

Melhem, E. R.; Hoon Jr, A. H.; Ferrucci Jr, J. T.; Quinn, C. B.; Reinhardt, E. M.; Demetrides, S. W.; Freeman, B. M. and Johnston, M. V. (**2000**). *Periventricular leukomalacia: relationship between lateral ventricular volume on brain MR images and severity of cognitive and motor impairment*, Radiology 214 : 199-204.

Sarı, E.; Sarı, S.; Akgün, V.; Özcan, E.; İnce, S.; Babacan, O.; Saldır, M.; Açıkel, C.; Başbozkurt, G.; Yeşilkaya, Ş. and others (**2015**). *Measures of ventricles and evans' index: from neonate to adolescent*, Pediatric neurosurgery 50 : 12-17.

Turner, A. H.; Greenspan, K. S. and van Erp, T. G. (**2016**). *Pallidum and lateral ventricle volume enlargement in autism spectrum disorder*, Psychiatry Research: Neuroimaging 252 : 40-45.

Zhang, L.; Tanno, R.; Xu, M.-C.; Jin, C.; Jacob, J.; Cicarrelli, O.; Barkhof, F. and Alexander, D. (**2020**). *Disentangling human error from ground truth in segmentation of medical images*, Advances in Neural Information Processing Systems 33 : 15750-15762.