Group members

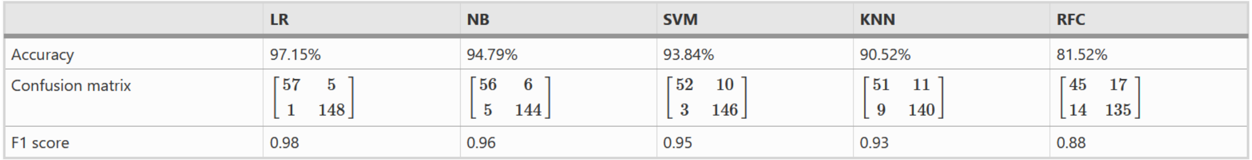
Name	Sec	BN
Habiba Mohamed Hanafy	1	25
Zeiad Ayman Mohammed	1	31
Omar Ayman Mohammed	2	2
Yara Hossam El-Din Mostafa	2	48

For the autism detection problem, we found various studies on the subject, but they had their own limitations for various reasons which include but not limited to:
1)unbalanced datase.
2)constrained sample size.
3)Missing instances from dataset that leads to overfitting.
There are some studies that used novel algorithms, but their results were less in accuracy than traditional ones or the increase in accuracy is subject to data variations.
reference:https://link.springer.com/article/10.1007/s42979-021-00776-5/tables/1

It's observed that most of the problems are encountered due to the dataset itself, which isn't the case in our project as the data set we are using doesn't have the aforementioned shortcomings.

For the next section, we will be referring to this study :
https://link.springer.com/article/10.1007/s42979-021-00776-5#ref-CR5

In the study, the same dataset we found is used, and a similar approach of choosing the ML models is observed. Five models were used which are: Logistic regression,SVM,Random forest, Naive bayes and KNN.
In the study, the same preprocessing procedure we suggested in phase 1 was used which is dropping the case no. column and who filled the test column.
Encoding of text data was also used so the data can be used in the models mentioned.
It was found that the logistic regression gave the highest accuracy in this study with 98% as shown in the table below.



	LR	NB	SVM	KNN	RFC
Accuracy	97.15%	94.79%	93.84%	90.52%	81.52%
Confusion matrix	$\begin{bmatrix} 57 & 5 \\ 1 & 148 \end{bmatrix}$	$\begin{bmatrix} 56 & 6 \\ 5 & 144 \end{bmatrix}$	$\begin{bmatrix} 52 & 10 \\ 3 & 146 \end{bmatrix}$	$\begin{bmatrix} 51 & 11 \\ 9 & 140 \end{bmatrix}$	$\begin{bmatrix} 45 & 17 \\ 14 & 135 \end{bmatrix}$
F1 score	0.98	0.96	0.95	0.93	0.88

Plan for the next phase:

We will continue with the suggested methods mentioned in phase 1 as it was found that the study used most if not all of the methods with a very high result.
We will focus on logistic regression as it gives higher accuracy than other methods.
We also may try to use preprocessing methods especially for logistic regression to see if we can raise the accuracy even higher.
Some columns which have text data will be encoded so we can use them in our models.