COMP0189: Applied Artificial Intelligence Coursework - 2

Surbhi Goel
Department of Computer Science
University College London

***Abstract*— The coursework aims to analyse and compare the effectiveness of low dimension and high dimension feature sets, machine learning models, and cross-validation strategies for predicting the presence of schizophrenia using brain grey matter data. The study focuses on a classification task, where the ability of the models to accurately classify patients as having schizophrenia or not is evaluated. By comparing the performance of different feature sets and models, this study seeks to identify the best approach for predicting schizophrenia using brain grey matter data. Different cross-validation strategies have been explored in the coursework to ensure the reliability and generalizability of the results. Overall, this coursework aims to contribute to the development of accurate and reliable diagnostic tools for schizophrenia.**

***Keywords*— schizophrenia, prediction, machine learning, cross-validation**

I. INTRODUCTION

Schizophrenia is a critical mental disorder that affects people worldwide. It is important that the disorder is diagnosed accurately and an effective treatment is employed timely. Brain imaging data has been of utmost importance for doctors and scientists to understand the physical changes that take place in the brain during schizophrenia. Grey matter in particular has been of utmost interest to predict schizophrenia. The grey matter data cane be divided into low dimension features which are the crudely extracted features of the grey matter and high dimensional features which are more detailed features. The aim of our coursework is to analysis the impact of these low and high dimension features in predicting if a person is schizophrenic or not. The features will be analysed through different models and cross-validation techniques. By identifying the best approach for predicting schizophrenia, the coursework aims to contribute to the development of accurate and reliable diagnostic tools for the schizophrenia. The coursework explores different strategies to ensure the reliability and generalizability of the results, providing valuable insights for future research.

II. METHODS

A. Comparison of Models

There are three machine learning models that have been compared and analysed to perform classification of schizophrenic and non-schizophrenic patients. The three models are, Logistic Regression (Linear Regularised model), Random Forest (Ensemble) and Support Vector Classifier (Non-linear model.

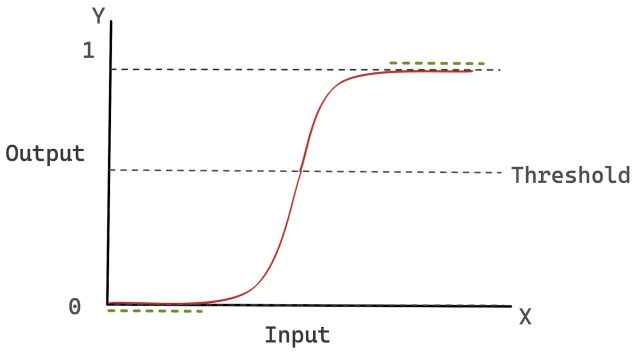

Fig. 1. Sigmoid Function of Logistic Regression

*1) **Logistic Regression (Linear Regularised model):*** A linear model is where the relationship between input and target variables is assumed to be linear. A regularized linear model when a regularization term is added to the objective function. The purpose of adding the regularization is to prevent overfitting by adding a penalty to the model's parameters so that alterations in the parameter values do not shoot and can be constrained within a range.

Regularization can be of various types:

- L1 or Lasso Regularization
- L2 or Ridge Regularization
- Elastinet

Logistic regression is a linear algorithm that principally uses a Sigmoid function to give probability of the target value, using a set of input features. A threshold value of probability can be set to classify the data.

It has the advantage of being a simple and interpretable linear model that can be used for both numerical and categorical input features. But when the relationship between input and output data is highly non-linear, the logistic regression model may not be able to capture the patterns accurately. As our coursework is a comparative study, using a simple logistic regression model is makes sense, as we can compare it's performance with other more complex algorithms.

The model used in our analysis has the following attributes for RBI and VBM data, using two different types of cross-validation and can be seen in the Table I and Table II.

The C-value is the strength of regularization of the

TABLE I
BEST PARAMETERS FOR LOGISTIC REGRESSION MODEL
(COMMON CROSS-VALIDATION)

	C-value	Regularization	Solver
ROI	0.1	L2	lbfgs
VBM	100	L2	lbfgs

TABLE II
BEST PARAMETERS FOR LOGISTIC REGRESSION MODEL
(GROUP-STRATIFIED CROSS-VALIDATION)

	C-value	Regularization	Solver
ROI	0.1	L2	lbfgs
VBM	10	L2	lbfgs

model and lower the value, higher the regularization strength. The best C-value for ROI data remains the same for both types of CV techniques i.e. C=0.1 which is a high strength regularization. But for VBM data, the best C-value is high for both types of CV techniques. Other parameters are kept fixed at Solver= *ibfgs* and Regularization= *L2(Ridge)*. L2 regularization has been used because L2 has a squared penalty component which provides smooth and more stable solution. It also helps prevent overfitting in data where multiple features have high level of correlation.

The models has been trained on the above parameters for ROI and VBM data and the trained models has been tested on test ROI, VBM data. The results have been compared in the Results section (fig 4 and fig5 respectively)

2) ***Random Forest (Ensemble)****:* Ensemble is a machine learning models that involve combining multiple learning models to improve the overall performance. These methods are also beneficial for quantifying uncertainty in predictions as models with different strengths capture the patterns over a wide range and can help achieve predictions with highest probability.

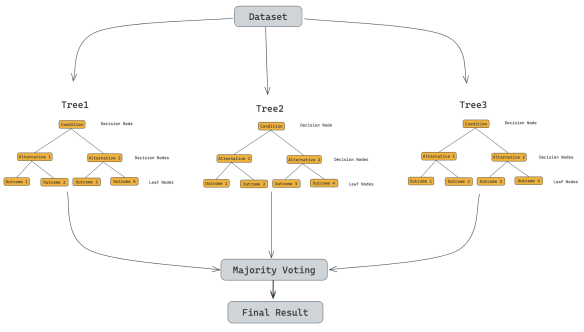

Fig. 2. General Architecture of Random Forest

The combinations of models categorises ensembling into three types,

- *Bagging*: Models are trained in parallel
- *Boosting*: Models are trained sequentially
- *Stacking*: Models are trained in a hierarchical manner

Random Forest is a type of bagging ensemble method, that combines multiple decision trees, to predict the outcome of a problem. Random Forest constructs a large number of decision trees and each of the tree is trained on a different subset of the data and features. The final prediction is made either by by averaging the predictions of all the individual trees or by majority voting among the trees.

This strategy improves the overall accuracy and provides a more generalized model. Random Forest Models are considered good for dealing with data that has non-linear relationship between input and target variables. It is a complex, yet interpretable model for machine learning.

The model used in our analysis has the following attributes,

TABLE III
BEST PARAMETERS FOR RANDOM FOREST MODEL
(COMMON CROSS-VALIDATION)

	max_depth	min_s_split	min_s_leaf	n_estim
ROI	20	4	10	50
VBM	10	1	5	20

TABLE IV
BEST PARAMETERS FOR RANDOM FOREST MODEL
(GROUP STRATIFIED CROSS-VALIDATION)

	max_depth	min_s_split	min_s_leaf	n_estim
ROI	10	4	50	50
VBM	20	1	50	20

Due to being computationally heavy, the range of no_of_estimators for ROI and VBM dataset have been taken different. ROI being low-dimensional data, took a couple of minutes, whereas tuning and training VBM data took enormous amount of time, as discussed in Results section ahead. The no_of_estimators for VBM dataset had to be reduced to keep the training time under an hour. Interestingly, the hyperparameter tuning algorithm chose the lowest available no_of_estimators for both ROI and VBM datasets.

As seen in the tables above, the best set of parameters achieved through hyperparameter tuning have been used to train the model. ROI and VBM test data has been used to calculate scores attained by the model using the best set of parameters. These results have been discussed in the Results section.

3) ***Support Vector Classifier (Non-linear model)*** *:* Support Vector Machines are non-linear models, i.e. models that assume that a non-linear relationship exists between input and target variables. The working principal of SVM is finding a hyperplane that separates the different classes

in the feature space. The input used is usually called as feature vector.

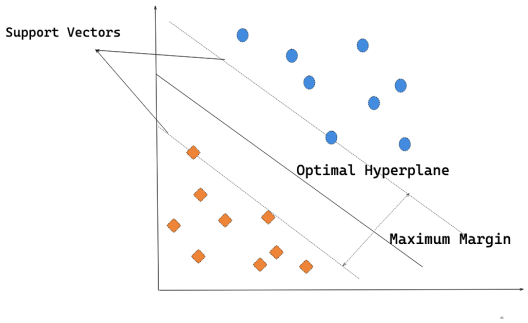

Fig. 3. Working Principal of Support Vector Machines

TABLE V
BEST PARAMETERS FOR SUPPORT VECTOR MODEL
(COMMON CROSS-VALIDATION)

	C-value	Kernel
ROI	10	rbf
VBM	0.001	linear

TABLE VI
BEST PARAMETERS FOR SUPPORT VECTOR MODEL
(GROUP STRATIFIED CROSS-VALIDATION)

	C-value	Kernel
ROI	100	rbf
VBM	10	rbf

SVC has been selected as the one of the models for comparison of performance on the Schizophrenia dataset because the dataset is complex and SVC generally performs well on complex dataset with large number of features. Although it is considered as computationally heavy because of the working principal of SVC that involves transforming low dimensional data to high dimensional. The above tables (Table V and Table VI) have the list of hyperparameters achieved through performing hyperparameter tuning. *rbf* has been a favourite choice of the tuning algorithm and is justified as *rbf* kernel can smoothly handle non-linear, large scale dataset.

B. Pipeline

Cross-Validation has been performed through two types of techniques, namely - *Common CV* and *StratifiedGroupKFold CV*.
Common Cross-Validation divides the data into K equally sized subsets and uses one subset for testing while the remaining subsets are used for training. For our analysis, we have kept k-value=5 (no. of subsets) for both ROI and VBM datasets.
StratifiedGroupKFold Cross-Validation divides the data into K folds and also takes into account the class distribution and group structure of the data. This method is usually considered better when data is unbalanced.
Both these cross-validation techniques have been compared for ROI and VBM data, for all three mentioned models.

For hyperparameter tuning, GridSearchCV algorithm by sklearn library has been used. The algorithm performs search over the parameter grid to find the best hyperparameters for a machine learning model. The tuning algorithm evaluates the performance of each combination of hyperparameters and selects the one with the best performance. For each of the model, a range of parameters have provided to the GridSearchCV algorithm and the parameters have been mentioned in the previous section.

The dataset was divided into ROI and VBM dataset as mentioned in the challenge, and both datasets were used to train and test the parameters selected by the hyperparameter tuning algorithm.

C. Metric

The choice of metric depends on the problem at hand. Since we are dealing with a healthcare data, where the aim is to classify patients as Schizophrenic based on certain features, the metric plays an important role to measure the scalability and reliability of the model.

For our dataset, it is important that the model does not give False Positives or False Negatives, because in both the cases, the patient may have to deal with great level of inconvenience and can even prove to be fatal. So, the performance metric chosen here is Recall and Precision. Recall helps to subside False Negatives and Precision helps with False Positives. So, the performance of all three models have been measured and compared using these metrics (Fig4, Fig5).

For Computational Cost, the total time taken for hyperparameter tuning and model training have been calculated and compared. It is a fair metric to check if the model is computationally feasible or not. The time has been collected in *minutes* and has been used to compare the computation of each of the test case (Fig 6 and Fig7).

III. RESULTS

The performance of all three models have been compared, using 2 levels of datasets, and cross-validated using 2 different techniques. The comparison has been done in term of model performance and computational cost. The test cases have been analysed in our study and the results have been plotted below, in Fig4, Fig5, Fig6, Fig7.

Fig 4 shows the comparison of performance of ROI and VBM features for common cross-validation. The three models (Logistic Regression, Support Vector Classifier and Random Forest) have been compared on the basis of their Precision, Recall and F1 Scores.

Fig5 shows the comparison of performance of ROI and VBM features for StratifiedGroupK-fold cross-validation. The same three models have been compared.

For computational cost comparison, the time taken for hyperparameter tuning and the total time taken for the model run has been plotted.

Fig6 shows the comparison of computational cost(in terms of time) of all the models. The plot shows the number of minutes taken by each of the model, with each of the feature set(ROI and VBM), for the common cross-validation technique.

Fig 7 shows the comparison of computational cost(in terms of time) of all the models for ROI and VBM, for Stratified Group K Fold Cross-validation.

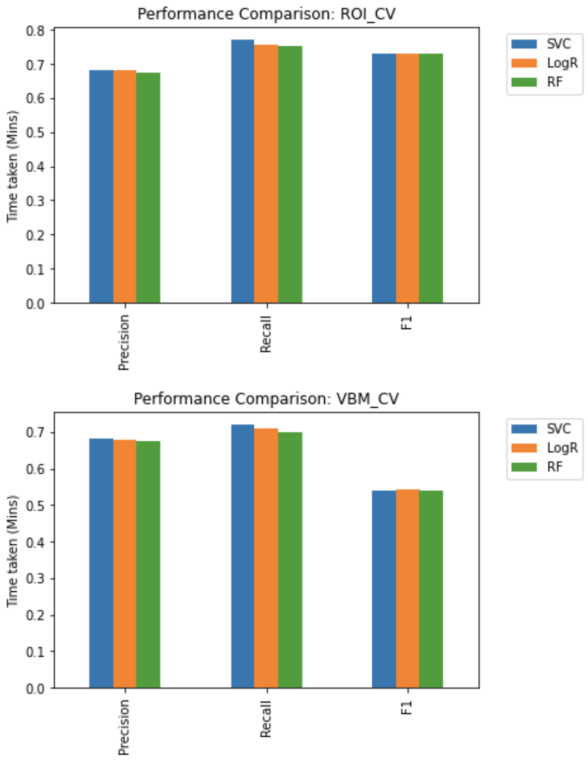

Fig. 4. **Performance comparison of models (Common CV)**

IV. Discussion

For the ROI (low-dimensional) features, all three models perform approximately the same, but the **Support Vector Classifier** seems to have a better Recall score, the model is able to correctly identify most of the instances that belong to the class of interest, and there are fewer false negatives. Even for Precision score, SVC performs better, i.e. the model is able to correctly identify a large proportion of the observations that it classified as positive, while minimising the number of false positives.

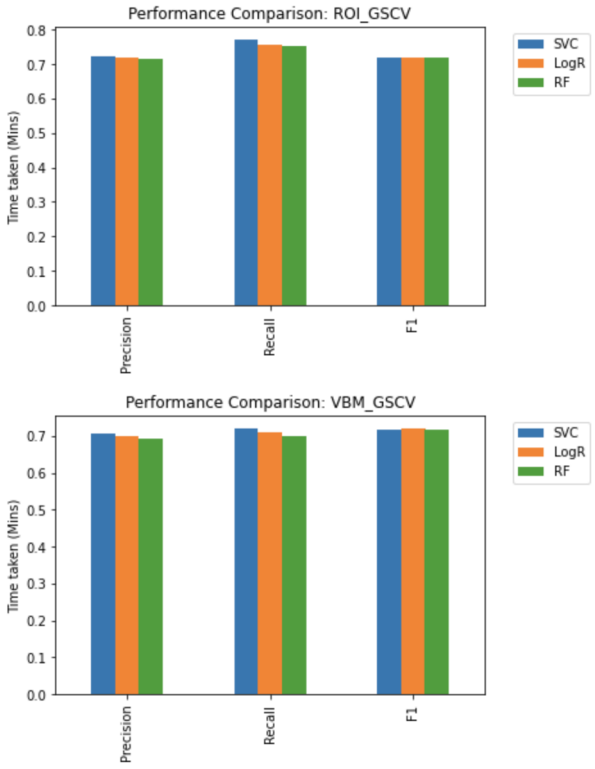

Fig. 5. **Performance comparison of models (Stratified Group CV)**

Logistic Regression Model is a not-so complex model, but it still managed to perform decently as compared to other models. Random Forest on the other hand, does not perform too well, as opposed to what ensembles are expected to.

For computational cost, **Logistic Regression** has been the least time-consuming model, followed by Support Vector Classifier. Random Forest has been slightly more time consuming, but still falling within few seconds.

For the VBM (high-dimensional) features, **Support Vector Classifier** performs better as compared to other models. It has a higher Recall and Precision score, which is highly desirable.

For computational cost, **Logistic Regression** has been the least time-consuming model, followed by Support Vector Classifier. Random Forest has been extremely time consuming (50 minutes).

The two cross-validation techniques have been compared in Fig 6 (Common CV) and Fig 7 (Stratified CV). If we look at the figures, we see that Stratified GroupK-Fold has achieved higher Recall, Precision and F1 Scores (Fig7) compared to Common CV (Fig6) for both ROI and VBM. Although the relative performance between models does not seem to alter much, but the overall improvement of results in case of Stratified CV is quite evident.

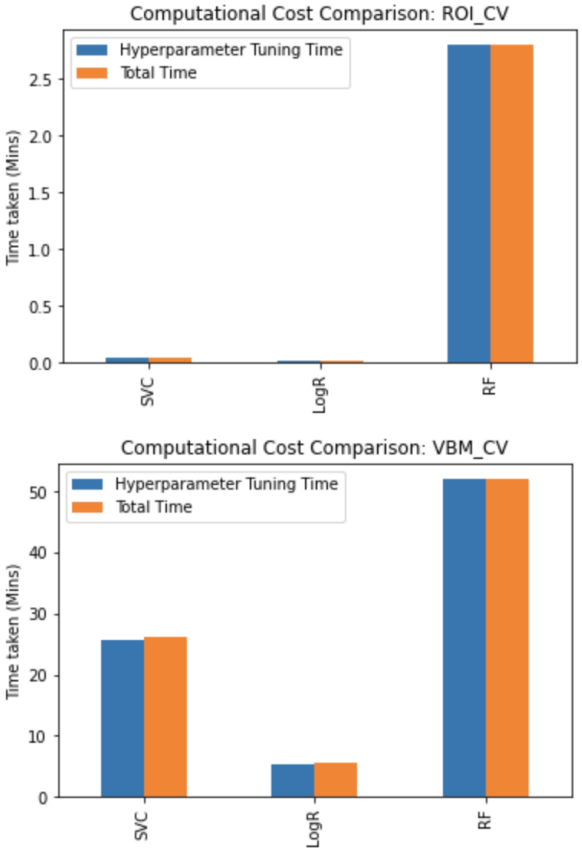

Fig. 6. **Time taken by models (Common CV)**

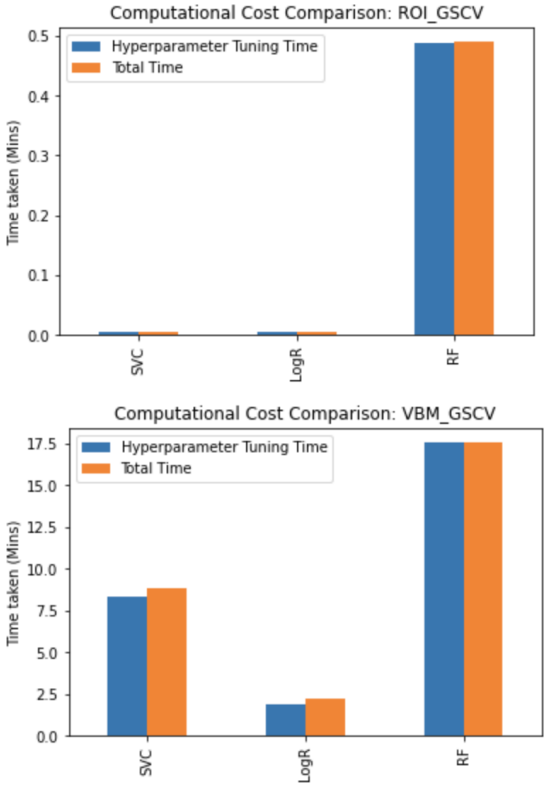

Fig. 7. **Time taken by models (Stratified Group CV)**

V. Conclusion

In conclusion, there were three models of different capabilities,

- Logistic Regression - Linear, simple
- Support Vector Classifier - Non-linear, complex
- Random Forest - Non-linear, super-complex

But if we look the plots, the prediction performance of the models has been approximately same, with no extreme differences. Although, this should not have been the case, as all three models had different strengths.

Logistic Regression being a simple model gave a Precision score of around 0.7 which is approx. equal to SVC and Random Forest. Computationally also, Logistic Regression performed extremely efficiently and took minimum possible time.

Random Forest unexpectedly performed poor and has extremely high computation cost when fed with large amount of high dimensional features (VBM).

Support Vector Classifier performed well prediction-wise as well as computationally.

Thus, it shows that depending on the data, simpler models can prove to be more useful than complex ones. A simple model with fewer parameters may be able to capture the underlying patterns in the data more effectively than a complex model with more parameters. This is because complex models can be prone to overfitting, which occurs when the model becomes too complex and starts to fit the noise in the data instead of the underlying patterns.

Also, adding more effective cross-validation techniques can certainly improve the overall performance of the machine learning models.