New Cytomine modules for user behavior analytics in digital pathology

Master Thesis

Academic year 2017-2018

University of Liège - Faculty of Applied Sciences

Graduation Studies conducted for obtaining the Master's degree in Computer Sciences by Laurent Vanhee

June 2018

Contents

1 Abstract

In the medical field, doctors and researchers need to be able to observe and interpret cell samples. The most widely spread method is to observe samples with a microscope. New methods allow us to scan a sample into a very large and detailed image. These images can be uploaded to the Cytomine web application and doctors can go through the images with ease. The application includes many functionalities, namely the ability to annotate regions of interest. Currently the ULiege MOOC server is used for the benefit of the students studying the medical field at the University of Liege. The students study particular images and are then evaluated at the end of the year. Meanwhile, the Cytomine app has been collecting data on the studentsâĂŹ time spent on the website. The bulk of the data collected consists of where the students decided to look in the images (Gaze data). With that, attempts were made to find correlations between studentsâĂŹ behavior and the results they obtain during exams. Using Machine Learning techniques, the goal is to predict a studentâĂŹs grade based on how they used the application. Currently, the model contains 395 students with over 2000 features. Random Forest and Extra Trees learning techniques have been applied to attempt to predict grades. Otherwise, another goal is to visualize these patterns. The idea would be to generate Heatmaps of the studentsâĂŹ gaze data (Gazemap). These Gazemaps would be included in the app and users can be given access to this information. This could give teachers the ability to keep track of the studentsâĂŹ work.

2 Introduction

2.1 Cytomine

Open-source rich Internet application for collaborative analysis of multi-gigapixel images. Cytomine can be described by three main properties :

- Open Source : The source code is available to the general public, in this case with Github. It has a Apache 2.0 License, which is very permissive for any third party. It is also accompanied by documentation that describes the different modules and how to use them.
- Open Company : It's a non-profit company that contribute and promote the project.
- Open Research : Cytomine employs researchers at the Montefiore Institute of the University of Liege. They develop Machine Learning algorithms, image informatics, and Big Data modules. Cytomine also collaborates with other researchers.

Cytomine was developed to ease the analysis of multi-gigapixel images. These images can take Gigabytes of disk space. For most computers, displaying such images at full resolution is impossible. With Cytomine, images are stored in a server. These images can be viewed using the web application with any modern browser. Cytomine handles everything locally so that performance is not an issue for the clients.

Cytomine also comes with a comprehensive and robust API (Application Programing Interface). This API allows clients to fetch and send data to and from a Cytomine Server.

This is very useful for data analysis, it's well organized so that researchers can find what they are looking for. There are also tools included that eases certain aspects of handling big data. For examples image annotations allows the user to create zones of interest. These zones can share characteristics and using learning algorithms, researchers can learn new zones of interest.

2.2 Cytomine for education

Cytomine can be used as a tool of education for many subjects. Since it's open source and well documented, it is possible to fork it and develop new methods and modules. This can be done for many fields in computer sciences including Machine Learning, Vision, Big Data, or even front-end programing.

This is not limited to the field of Computer Sciences. Fields that require the use of large high definition images can benefit from using Cytomine. This includes astrology where, people can learn about planets and stars using imagery. Other sectors include Geology, Art, and in this case Medicine.

2.3 MOOC Server

The MOOC server is a Cytomine Web Application used for education by the Faculty of Medicine at the University of Liege. It is used for the HISL0541 (General histology and alternative experimentation methods that do not use animals) course. This course is organized for students enrolled for a Bachelor in Medicine. Students have to study tissue samples.

Cytomine is a great tool for this particular field because it can be used to store and display high quality images of tissue samples. It is also easily accessible for the students, they are not restricted to only using it during lab sessions. It's available 24 hours a day, therefore it gives students a good amount of liberty. The details of the assignments are given by Ecampus which is communication tool for students and teachers delivered by the University. The teachers can also use Cytomine to explain certain visual concepts in real-time. They can also use certain images for exercises by telling students to find patterns in those images.

Since the MOOC is completely integrated to the course, it directly impacts how students learn. Their exams are often based on what they learned using the tool. The goal is to see how much of an effect Cytomine has on students.

2.4 Teachers' Opinion of the Tool

2.5 General Goals

3 Tools and Methods

3.1 Three Separate Modules

Since there is a vast amount of data, most operations require some good amount of time to complete. Therefore, the data analysis tool is divided into three modules :

- Data Acquisition.
- Data Manipulation.
- Data Learning.

3.1.1 Data Acquisition

During the 2016-2017 Academic year, Cytomine tracked and stored user information on the MOOC server. To obtain all this information, Cytomine is accessible by a REST API. To ease the access to the server, Cytomine developed a Python Client that was used for the acquisition of relevant information. With administrator rights, a user can get a hold of all if not most of the data stored on the SQL and MongoDB databases.

There are currently a total of two projects called GOLD and SILVER that students could participate in. To start of, each project contains a set of images and a set of students that have signed up.

	Number Of students	Number Of Images
GOLD	395	78
SILVER	85	75

Both projects have their own objectives but GOLD is more complete and thorough. unlike the SILVER project, the students were tested and graded on the course that was given to them but also the content of the project.

To analyze user behavior, it is necessary to fetch data that is relevant, this includes :

- Resized images :

  The images stored on Cytomine are in fact very large. For the analysis, a copy of the images will be useful. Many operations in the Data Manipulation module rely on the image resolution when it comes to complexity. It is important that the image is big enough to be viewable while small enough so that the operations done in a reasonable amount of time. The image downloaded is therefore rescaled to a maximum width and height of 1024 pixels.

- Reference Annotations :

  When observing images, students are usually given guidelines in forms of image annotations. These annotations are zones in the image that contain information that students can learn from. These annotations are given a number. Annotations from the same image have a different numbers. This represents the recommended order the user can traverse the image. In this study, only the geometrical center of the annotation is kept. In most cases, this loss of information should have no impact because the size of the zones are usually a couple pixels wide when put in the resized image.

- User Annotations :

  Teachers can set annotations as guidelines, but normal users can also create annotations. If a student user notices something interesting on a patch of an image, that student can annotate it. Later, that student could for example approach a teacher with a question and use the annotation as a reference. Unfortunately, there are currently no User annotations. This will be discussed in section 5.

- User Positions :

  The most important information. A Positions is what the user sees at a current time stamp. Positions are defined by its center, four corners, time recorded, and zoom. Positions are saved on a regular basis when a user observes an image. More precisely positions are saved :

  - Every 5 seconds.
  - When the user switches zooms.
  - After the user finishes a movement on the image.

  Due to how frequently positions are recorded, this information comes in large quantity.

- Annotations Actions :

  Annotations are clickable. When clicked, a toolbox appears giving more information on the annotation. This action is also stored on the server. For the data recorded in 2017, a annotation action only contains a time stamp. It is only in later versions of Cytomine that the reference annotation identifiers were tracked with the annotation actions. In the case where the referenced annotation is unknown, it will be guessed based on positions that appear at the same time.

All this information needs to be downloaded. Unfortunately, this can take up to 8 hours using the API while putting stress on the MOOC server. After some interruptions to the service, it was decided that the MOOC needed to be installed locally with backups of the

original.

For both projects, an excel file containing user information not found on Cytomine was given. For the SILVER project, this only included basic information that were irrelevant to the analysis (names, emails, etc..). Meanwhile for the GOLD project, the University of Liege students had to partake in multiple graded assignments. The excel file given for the GOLD students therefore contained grades for numerous activities. These grades will be useful in order to understand some aspects of the user behavior on Cytomine.

The project has a dedicated directory to store all this information, usually in a csv format. For each data type, image, and user triplet, there is a dedicated file containing this information. These files will be opened and read by the Data Manipulation Module for generating statistics and ways to represent data visually.

3.1.2 Data Manipulations

This module is used to interpret the data that's in its most basic form. It has multiple outputs based on the parameters set.

3.1.3 Data Learning

3.2

4 Data Analysis

4.1 Experiments

Even though the data was obtained for both the GOLD and SILVER projects, the experiments were on run on the GOLD project. This is due to the many constraints given by the SILVER project including the sample size and the lack of results (student grades). The goal for most experiments is to learn a specific grade the student obtained based on all the information gathered in the Data Manipulation module. There were over 2000 features generated, where each one can weigh in on the prediction. These features were used to learn over 13 grades including the final grade obtained by these students (first session).

Due to the nature of the dataset, regression trees were the most ideal. The somewhat small dataset paired with a large amount of features is a big constraint to work with. Therefore, Ensemble methods were used and tested. This includes Random Forest regression and Extra Trees regression using sklearn.

The Learning module is given a statistics file containing rows of students paired with their features. The features are split into three categories :

- **M** : Meta-data variables, these variables have no statistical significance. These variables usually contain basic information on the users.

- **X** : Variables with statistical significance, these variables mostly include data extracted from the Cytomine website. Either individual image variables or variables on the set of images. These variables are used in the machine learning model as input variables.

- **Y** : Result variables, these variables is what the algorithm is attempting to guess using the **X** variables. These variables are used in the machine learning model as output variables.

4.2 Data Set

4.2.1 Students

Like mentioned earlier, there are a total of 395 students in the data set. Students are defined by their features and the grades obtained. These student followed the course HISL054, "General histology and alternative experimentation methods that do not use animals" at the University of Liege during the academic year 2016-2017. Most of the work done by the students is done online using the MOOC and Ecampus. Ecampus is a website used by students and teachers of the University of Liege to exchange information used for courses. In fact all of the assignments for this course are explained on Ecampus. The issue with this is that it's a different entity from the MOOC and therefore is not ideal for fetching information. But it's not much of a problem because there's nothing to retrieve about students on Ecampus that is not already known.

When applying machine learning techniques, students with a 0/20 were taken out of the sample. This is due to the fact those students defaulted to signing the exam. Since there are too few that do so, there's not enough information to find correlations between people who sign and the features. Keeping students who sign also increases the error rate because they are extreme cases. The goal is to guess how well the exam would have went for the students. If the student decides not to bother with taking the exam, it's a lost sample.

4.2.2 Teacher Input : Grades

The students are evaluated by multiple exams and quizzes, these take the form of:

- QCM : Multiple choice question test. Students are given questions and have to respond with 1 out of the 9 (or less) options. They have to set a degree of certainty for each response. The exam results are then calculated by a machine. The students are also given more boxes to answer in case they make a mistake the first time.(Figure 4.1)

Figure 4.1: QCM sheet and question example

- QCL : Identification and Incidence test. Graded similarly to a QCM, the students are given multiple questions to answer and they have to fill out a form. The are a couple differences. Each question is split into two, the identification and the incidence. For the identification, the students have to identify an object on a image. They are given a exhaustive list of possible answers ranging from cells to tissues. Each answer contains is associated to a 3 digit code that they need to write on the form. For the incidence, the students needs to identify how the observed object "was cut". There are a total of 3 possible answers, transversal, longitudinal, and undetermined. The answer is written on the form under the identification answer.(Figure 4.2)

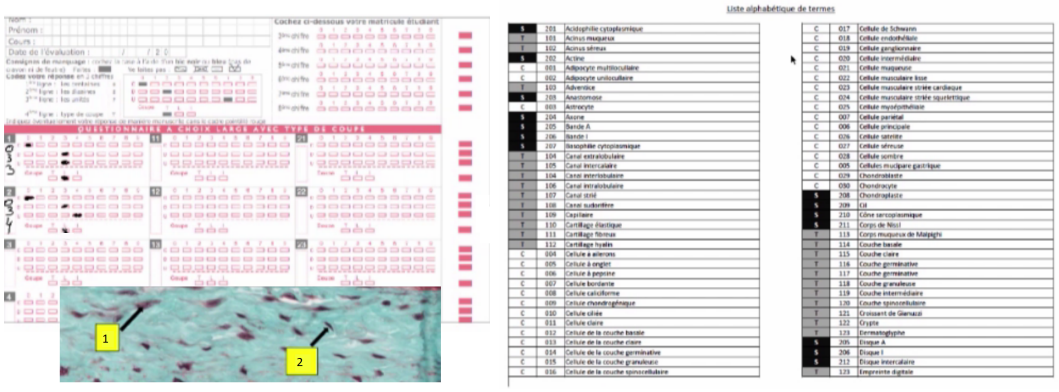


Figure 4.2: QCL sheet with question example and list of answers

- QROL : Long answer open question test. Students are to write and explain their answers in a detailed fashion. (Figure 4.3)

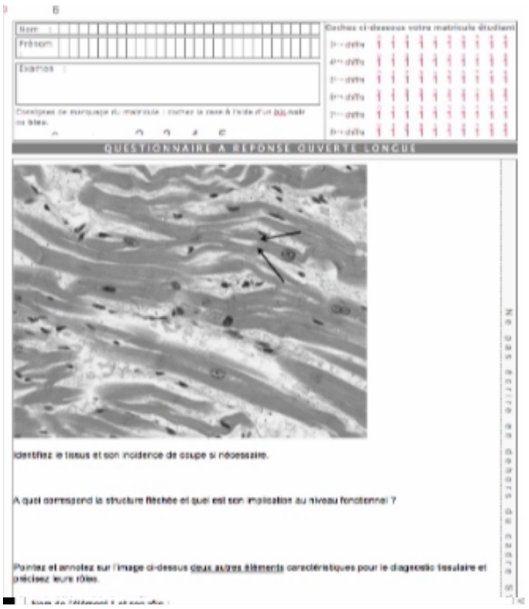


Figure 4.3: QROL sheet with a set of questions

These variables will be denoted as **Y** variables for output. Out of the 13 results associated to the students, 3 were white tests given as a practice tool :

- QCL identification white test.
- QCL incidence white test.
- Practical QCM white test.

Similarly, 7 graded exams and quizzes were given to students with 3 being theoretical and 4 being practical. Something to note for the practical exam is that there are 2 different exam forms for the QCM and the QRL. This means that half the students are given different questions from the other half. The list of exams include:

- QROL1 theory (10%).
- QROL2 theory (10%).
- QCM theory (30%).
- QCM practical (20%).
- QROL practical (10%).
- QCL identification Practical (16%).
- QCL incidence Practical (4%).

Finally, based on the previous results given, there are:

- Total Theory (50%)

- Total Practical (50%)
- global Grade (100%)

Most of the experiments are done using one of these grades as Y variables. In later experiments, some variables will be set as features. The learning of final grades using white test results as bonus features (**X** variable) could yield better results. Learning theoretical grades while also using practical grades and vice versa can also give some interesting results.

Teachers also input basic student information, but it mostly consists of general information that won't be used in the experiments. These will be denoted as **M** variables:

- ID Cytomine : The Cytomine ID associated to the user. (Compulsory)
- LAST NAME : User's last name.
- FIRST NAME : User's First name (forename).
- GROUP : group user belongs in (GOLDULiege, GOLD, or SILVER). GOLDULIEGE is a subset of GOLD containing the set of students following the course.
- USERNAME CYTOMINE : user's Cytomine username.

4.2.3 Features pre-Calculated

There are over two thousand features calculated and generated by the Data Manipulation module. These variables belong to the **X** category and are listed:

- NB IMAGES VISITED : The total number of different images that a user has opened over the course of the year.
- TOTAL NB POSITIONS : The total number of positions obtained from all the images opened.
- AVG NB POSITIONS : The mean number of positions obtained relative to all the images opened.
- MEDIAN NB POSITIONS : The median number of positions obtained relative to all the images opened.
- TOTAL IMAGE VIEWING TIME (s) : Total amount of time spent viewing images.
- AVG IMAGE VIEWING TIME (s) : Mean amount of time spent viewing images.

- MEDIAN IMAGE VIEWING TIME (S) : Median amount of time spent viewing images.

- NB POSITIONS AT ZOOM <x> : with <x> between 1 and 10, represents the zoom level of a position. 1 variable per zoom value. It represents the total number of positions at zoom <x>.

- AVG ZOOM : The mean zoom level over all the positions collected.

- MEDIAN ZOOM : The median zoom level over all the positions collected.

- TOTAL NB ANNOTATION ACTIONS : The total number of times the user clicked on a reference annotation.

- AVG NB ANNOTATION ACTIONS : The mean number of times the user clicked on a reference annotation.

- MEDIAN NB ANNOTATION ACTIONS : The median number of times the user clicked on a reference annotation.

- AVG NB POSITIONS AT ZOOM <x> : with <x> between 1 and 10, represents the zoom level of a position. 1 variable per zoom value. It represents the mean number of positions at zoom <x> relative to all images visited.

- MEDIAN POSITIONS AT ZOOM <x> : with <x> between 1 and 10, represents the zoom level of a position. 1 variable per zoom value. It represents the median number of positions at zoom <x> relative to all images visited.

- SCORE OF ANNOTATION <y> AT IMAGE <x> : with <y> being the annotation identifier and <x> being the image identifier. When images have annotations, students tend to focus on these points. For each position, the program generates a 2 Dimensional Gaussian function to represent the position. The size and values of this Gaussian relies on the zoom of the position. For example at the center of the Gaussian with the highest zoom, the value is 1. While at the lowest zoom the value is $1/MAX\_ZOOM$.

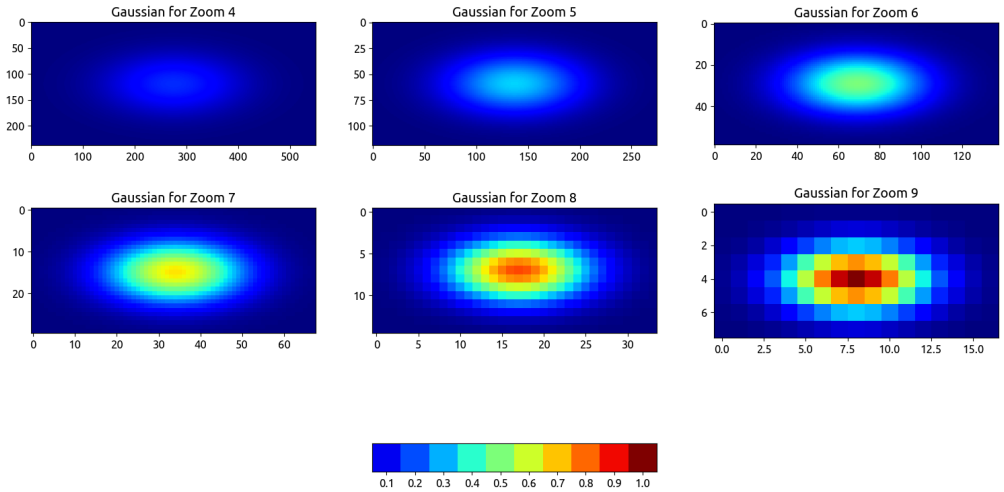


Figure 4.4: Example Gaussian Grids for a Image that has a zoom up to 9

So each annotation has a center coordinate. For this coordinate, a list of Gaussian values is generated based on all the positions near the annotation. This vector is sorted inversely. It's important to note that it does not take much time for the student to assimilate all the information on a part of a image. So the idea is to weight these positions to the point that after enough positions, the next ones would have a weight of close to 0. To do so, a geometrical sequence was used:

$$\sum_{i=0}^{N} w^i$$

N is the number of positions. In our case, w was set as 0.95. This means that this sequence converges to 20, and after about 50 positions the sequence is close to the convergence value (about 18). Since the list of Gaussian values are sorted, the highest values will have the highest weight. The equation becomes :

$$score = \sum_{i=0}^{N} L[i] * w^i$$

With L being the list of Gaussian values for the annotation. This means that the highest value has a weight of 1, the second a weight of 0.95, the third $0.95 * 0.95$, and so on. The value calculated is then used for the score. If a student actually observed in detail the annotation, they would usually get values above 10.

- USER SCORE AT IMAGE <x> : With <x> the image identifier on Cytomine. The users are given scores between 0 and 1 (not opening an image gives the user a score of 0). The score represents how well the user observed the annotations in the given image. In short, if a user spends a good amount of time on top of a annotation or if the user clicks on the annotation, he/she will get points for that annotation. Doing so for all annotations gives a final score. The score for an annotation is given by the previous variable (SCORE OF ANNOTATION <y> AT IMAGE <x>) and re dimensioned so this variable returns a score between 0 and 1. If the image does not have any annotation, the scores for are calculated for the entire image and the average is returned.

- AVERAGE USER SCORE : The average of all the scores defined previously for a user.

- NB POSITIONS AT IMAGE <x> : with <x> being the image Identifier, represents the number of positions recorded at that image for a user.

- TIME SPENT AT IMAGE <x> : With <x> being the image identifier, represents the total time spent on a image for a user.

- NB OF ANNOTATION ACTIONS AT IMAGE <x> : With <x> being the image identifier, represents the number of annotation Actions at that image for a user.

- NB OF POSITIONS WITH ZOOM <y> AT IMAGE <x> : With <x> being the image identifier and <y> being the zoom value [1-10]. It Represents the number of positions for a certain zoom at that image for a user.

- NB IMAGES VISITED DURING MODULE <x>: The total number of different images that a user has opened that are associated to the module <x>.

- AVG NB POSITIONS DURING MODULE <x>: The mean number of positions obtained relative to all the images opened that are associated to the module <x> during the corresponding time period.

- MEDIAN NB POSITIONS DURING MODULE <x>: The median number of positions obtained relative to all the images opened that are associated to the module <x> during the corresponding time period.

- TOTAL NB POSITIONS DURING MODULE <x>: The total number of positions obtained from all the images associated to the module <x> during the corresponding time period.

- TOTAL TIME SPENT DURING MODULE <x> (s) : Total amount of time spent viewing images associated to the module <x> during its given time period.

- AVG TIME SPENT DURING MODULE <x> (s) : Mean amount of time spent viewing images associated to the module <x> during its given time period.

- MEDIAN TIME SPENT DURING MODULE <x> (s) : median amount of time spent viewing images associated to the module <x> during its given time period.

- NB POSITIONS DURING MODULE <y> FOR IMAGE <x> : with <x> being the image Identifier of a image associated to the module <y>. This represents the number of positions recorded at that image for a user during the module's time period.

- TIME SPENT DURING MODULE <y> FOR IMAGE <x> : with <x> being the image Identifier of a image associated to the module <y>. This represents the time spent at that image for a user during the module's time period.

- NB ANNOTATION ACTIONS DURING MODULE <y> FOR IMAGE <x> : with <x> being the image Identifier of a image associated to the module <y>. This represents the number of annotation actions recorded at that image for a user during the module's time period.

- NB POSITIONS WITH ZOOM <z> DURING MODULE <y> AT IMAGE <x> : with <x> being the image Identifier of a image associated to the module <y>. This represents the number of positions recorded at that image for a user during the module's time period with zoom <z>. The zoom value <z> ranges from 1 to 10.

- AVERAGE ZOOM DURING MODULE <y> FOR IMAGE <x> : with <x> being the image Identifier of a image associated to the module <y>. This represents the average zoom level of all the positions recorded at that image for a user during the module's time period.

- MEDIAN ZOOM DURING MODULE <y> FOR IMAGE <x> : with <x> being the image Identifier of a image associated to the module <y>. This represents the median zoom level of all the positions recorded at that image for a user during the module's time period.

- NB POSITIONS AT ZOOM <z> DURING MODULE <x> : with <z> between 1 and 10, calculated for all the images associated to the module <x>. This represents the number of positions for a specific zoom for a user during the module's time period.

- AVERAGE NB POSITIONS AT ZOOM <z> DURING MODULE <x> : with <z> between 1 and 10, calculated for all the images associated to the module <x>. This represents the mean number of positions for a specific zoom for a user during the module's time period.

- MEDIAN NB POSITIONS AT ZOOM <z> DURING MODULE <x> : with <z> between 1 and 10, calculated for all the images associated to the module <x>. This represents the median number of positions for a specific zoom for a user during the module's time period.

- TOTAL NB ANNOTATION ACTIONS DURING MODULE <x> : With <x> being the module identifier, it represents the total number of annotation actions for a user during the module's time period.

- AVG NB ANNOTATION ACTIONS DURING MODULE <x> : With <x> being the module identifier, it represents the mean number of annotation actions for a user during the module's time period.

- MEDIAN NB ANNOTATION ACTIONS DURING MODULE <x> : With <x> being the module identifier, it represents the median number of annotation actions for a user during the module's time period.

- AVERAGE USER SCORE DURING MODULE <x> : With <x> being the module identifier, it represents the average predefined score for a user for the module's images during the respective time period.

- USER SCORE AT IMAGE <y> DURING MODULE <x> : With <x> being the module identifier and <y> the image identifier, it represents the predefined score at the image for a user during the module's time period.

- SCORE OF ANNOTATION <z> AT IMAGE <y> DURING MODULE <x> : With <x> being the module identifier and <y> the image identifier, it represents the predefined score of the annotation <z> at the image for a user during the module's time period.

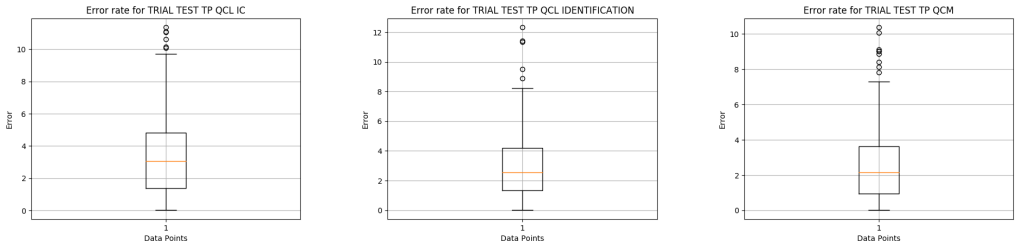
- PERCENT TIME WORKED AT NIGHT : Value from 0 to 1. This represents the ratio of the time spent working from 6pm to 6am.
- PERCENT TIME WORKED LATE : Value from 0 to 1. This represents the ratio of the time spent working from 1am to 6am.
- PERCENT TIME WORKED MORNING : Value from 0 to 1. This represents the ratio of the time spent working from 6am to 12pm.

- NUMBER OF DAYS WORKED : The total number of days where a user opened at least one image.
- PERCENT TIME WORKED DURING MODULE <x> : When working on Cytomine, the user can open images associated during a module during its time period or outside of it. This represents the ratio of the user activity during the time period.
- ANNOTATION <y> VISITED BEFORE ANNOTATION <y + 1> AT IMAGE <X> : This binary variable determines whether or not a student visited the annotation <y> before the annotation <y + 1> for a image <x>. The users are encouraged to study the reference annotation in a specific order.

4.3 Results

4.3.1 White Test Grades

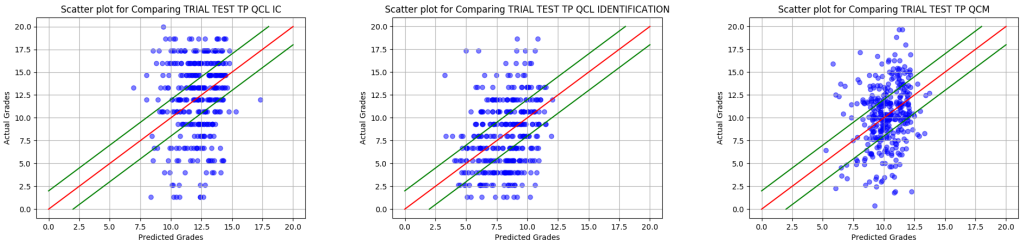
These 3 ungraded exam were taken halfway during the semester (24th of April 2017). They reflect what a student has learned and remembered up until that day. At that point, students should have finished modules 1 through 4 and at least looked at modules 5 and 6. These tests could be a great indicator on how well students can perform during the real exam. These **Y** variables were tested using a Leave-one-out cross validation on a Extra Tree Regressor with the generated data set. As a reminder, scores are calculated with the Mean Absolute Error technique. Also, the results are also compared with a model that learns by simply using the median. This is to determine if the model has much merit. (Figure 4.5)



4.5.1: Boxplots of the error values

	Score	Median Score	Score Difference	Real Median Grade	Median Est. Grade
QCL incidence white test	3.44	3.37	-0.08	13.33	12.321765
QCL identification white test	2.99	3.10	0.11	8.0	8.36
Practical QCM white test	2.58	2.57	-0.004	10.36	10.29

4.5.2: Discrete Results



4.5.3: Actual grades compared to Predicated grades

Figure 4.5: Results from cross-validation of the White Tests

Unfortunately, the results are not great. The boxplots show very high median and quartiles. There are also a high number of extreme errors with some above 10. The scores represent the average error of the cross validation. With scores nearing nearing 3, in some cases, it's better to try and guess using the median value. For the incidence test, it seems that the scores are underestimated, this seems to be the case because there is a significantly high grades compared to low grades. There are many reasons why results are the way they are. This includes the fact that these grades do not rely on activity that occur after the exam date. When comparing grade the learned grades against the original grades, it's noticeable that the estimated grades tend to predict values close to the average of the actual grades. This makes it so that high and low actual grades when predicted tend to be more erroneous. The sample also lacks a good amount of extreme grades which gives the algorithm less to work with for these cases. Furthermore, the fact that the exams are ungraded puts less pressure on the students to succeed. Therefore, they might not take the tests as seriously as they could by not reviewing and studying the days before.

Since the models are made using regression trees, it it possible to study the features that had an impact in determining the predicted grades. Even though the results were not ideal, it's interesting to observe what images and variables had a big impact on determining the grades. These images can very well be the subject of a question on the exam. (Figures 4.6,

4.7, 4.8)

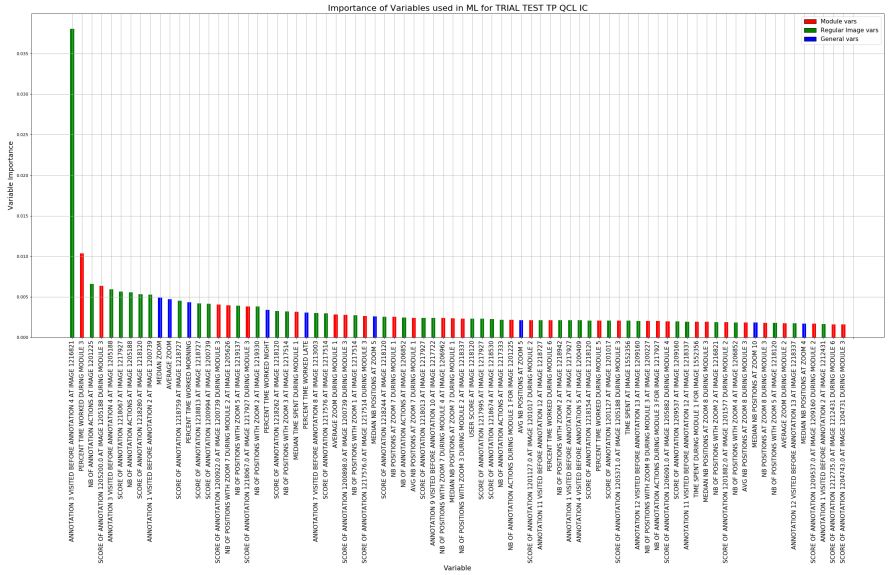


Figure 4.6: Feature Importance for QCL the Incidence

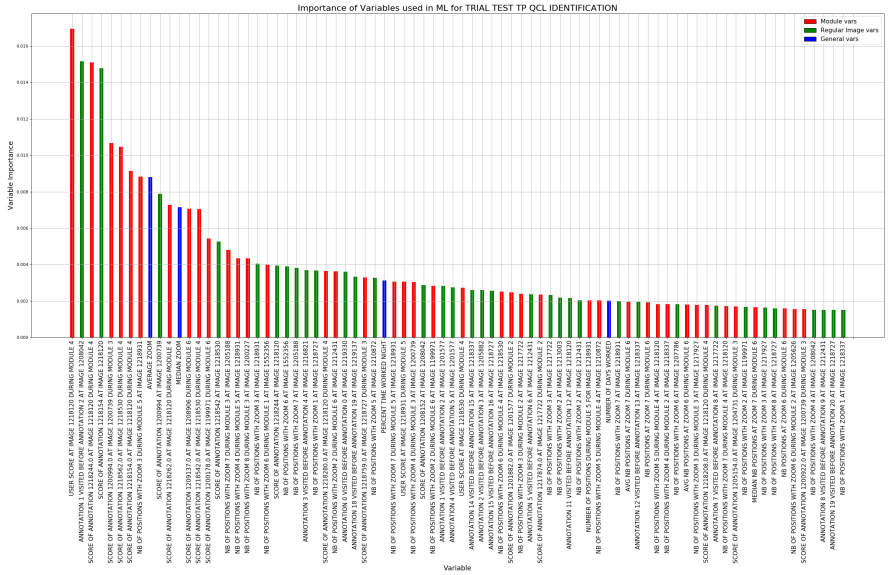


Figure 4.7: Feature Importance for QCL the Identification

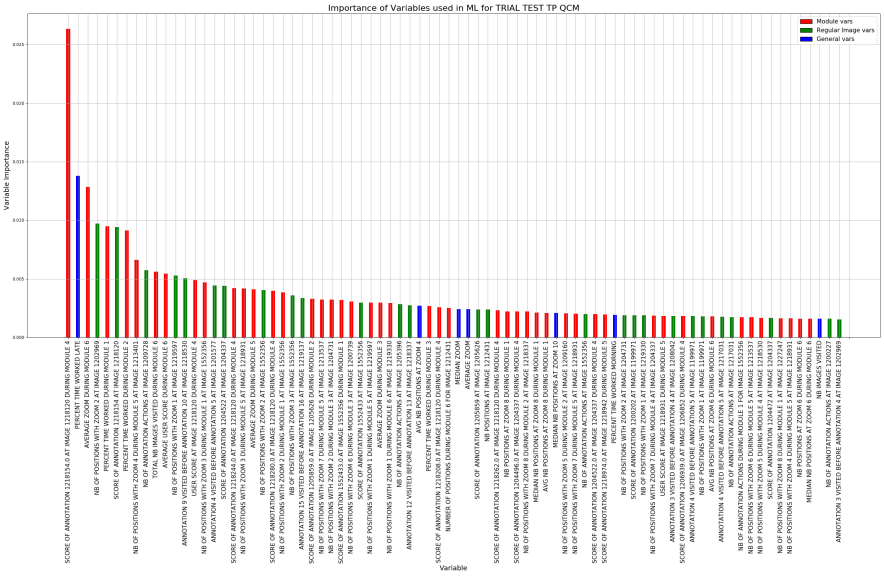


Figure 4.8: Feature Importance for the QCM

These figures show the top 80 features (or **X** variable) for their respective model. Even though each model shares the same features, their importance varies model to model. But it's probable that some features from a specific image can be at the top for multiple model. Anyways, features are split into 3 categories:

- General Features : Features that describes the data on the set of images as a whole. (average, median, etc..)
- Regular Image Features : Features that describe the data on a specific image. (number of positions at image XXXX)
- Module Image Features : Features similar to the two previous but associated to a specific module period. (number of positions at image XXXX during module Y)

Note that there are fewer general features, but many of those tend to have a high importance. Apart from the incidence test, it's appropriate to say that the module variables are the most impactful. This is because the white exam is taken in April instead of June. Therefore, user performance during specific modules are better indications on the performance during the white test. Also, for the identification test, the most impactful variable has a much higher importance than the rest. Even though this may look random, this variable may well be a good splitting factor. Unfortunately, this does not show the overall impact of certain images. (Figures 4.9, 4.10, 4.11)

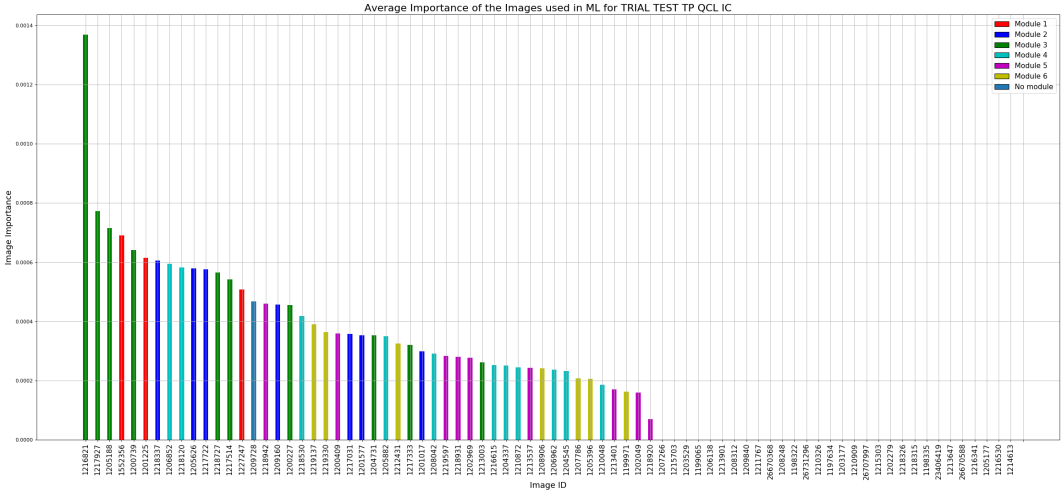


Figure 4.9: Image Importance for QCL the Incidence

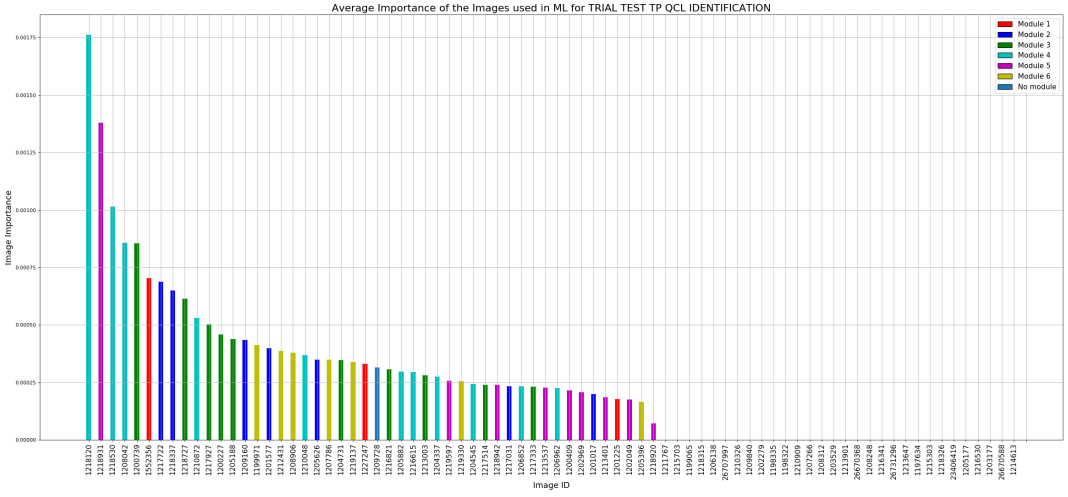


Figure 4.10: Image Importance for the QCL Identification

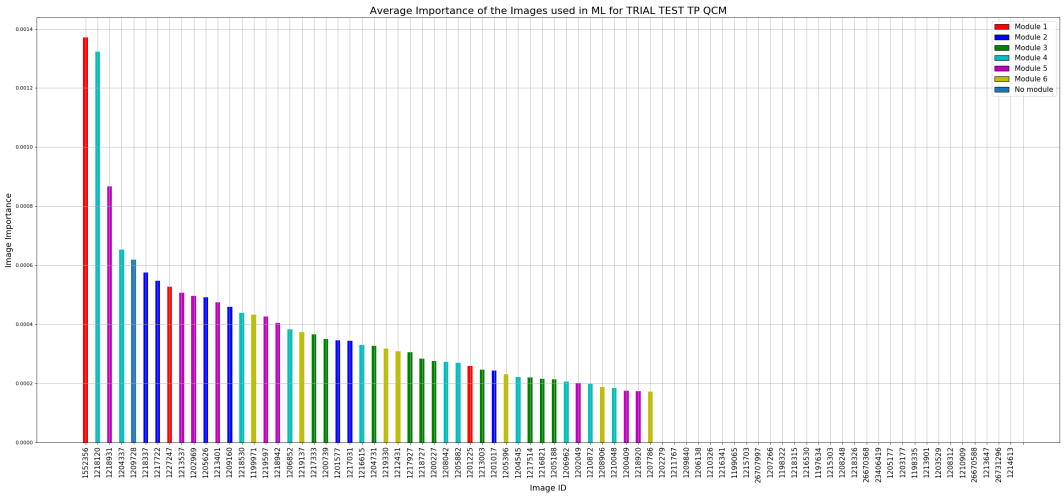
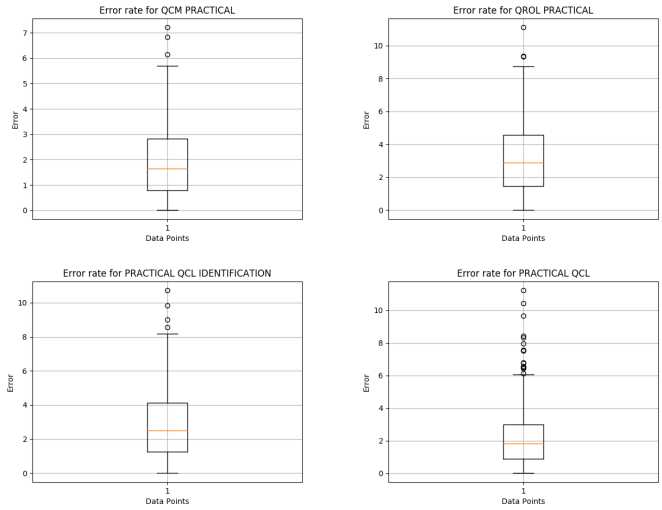


Figure 4.11: Image Importance for the QCM

As a reminder, out of all the images in the GOLD project, about a third of the images were not even visited. Also, some of the images don't belong to any module. It seems that as a general consensus that the most impactful images do not belong to the last module. This is due to the fact that the module 6 is not included in the trial exam. It's also worth noting that the module 1 is not included in the exam either. But since the module 1 is considered a tutorial, it can be useful. A image that appears often at the top is the image 1218120. This image was used for a preparation identification exercise. It seems that those who perform better in regards to that image on Cytomine do better in the white tests. For the QCL identification, the exercise images are often the most important over all the images in the same module. This seems to prove that doing the exercises seriously and correctly can lead to better grades. For the QCM, the image 1552356 which is a module 1 tutorial image is the most impactful. Those who follow the tutorial may have a easier time understanding the course from the start and may perform better for the QCM. As for the incidence test, the module 3 variables seem to be the most impactful. This module may contain examples that allow the student to better differentiate different incidence angles for specific objects in the images. OF course the cross validation results are not the best, so these assumption are not necessary correct. This leads to the actual practical exam.

4.3.2 Practical Exam Grades

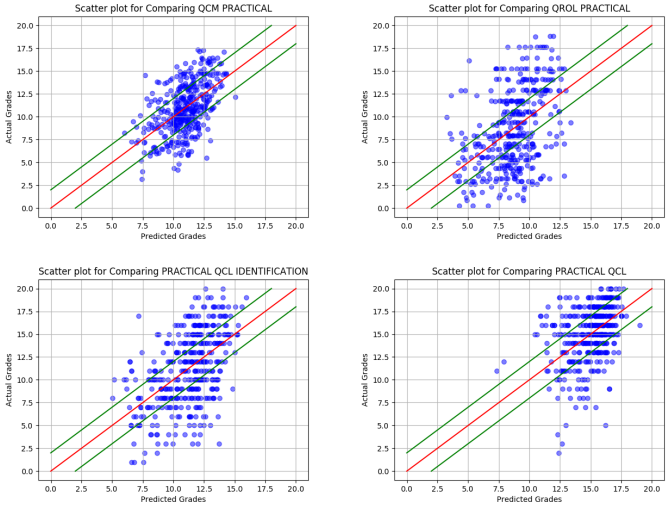
There are a total of 5 practical exam results including the total. These **Y** variables tested similarly to the white tests. The exam took place on the 12th of June 2017. Unlike the white exam, the students should have finished working on all the modules. These exams also depend on the entirety of the program unlike the white tests. Since these exams are graded, students should have the motivation to study and apply themselves. (Figure 4.12)



4.12.1: Boxplots of the error values

	Score	Median Score	Score Difference	Real Median Grade	Median Est. Grade
QCM practical	1.94	2.29	0.36	10.80	11.01
QROL practical	3.23	3.64	0.40	7.93	8.80
QCL identification Practical	2.86	3.36	0.5	11.0	11.61
QCL incidence Practical	2.23	2.83	0.59	15.0	15.32

4.12.2: Discrete Results



4.12.3: Actual grades compared to Predicated grades

Figure 4.12: Results from cross-validation of the White Tests

the results obtained are more interesting. The results vary heavily depending on the exams that were taken. All the scores are significantly better than their respective median compared to the trial tests.

- QCM practical : The best score, the boxplots show the most well rounded results. The 75 percentile of grades estimated have a error of less than 3 and a 25 percentile of grades with a error of less than 1. Unfortunately, there are still some cases with a significantly high error. As for relation between the expected grade and the estimated grade, there is a clear pattern following the red line. The exception being that low actual grades are overestimated by the model. QCM exams are usually are straight to the point. Students need to know the answer to question but they don't have to explain it. All students by definition are graded equally because the exam is straightforward and easy to correct.

- QROL practical : As opposed to the QCM, the score is significantly worse. This is also shown by the boxplot where the 75 percentile is at less than 4.5. By comparing the grades, it's apparent that there are more extreme grades and their estimations are off for the most part. This is normal because students write long and detailed responses to the questions. The teachers are more critical of the students. They look to see if students understand the contents of the course as opposed to learning by heart.

- QCL identification Practical : With a score of 2.86, this QCL does not offer the best results. It follows a somewhat lessened pattern of the QROL. As this portion of the exam gives the student a exhaustive list of possible answers, it's hard to a student to guess. It's critical that the students knows what they observe. Similarly to the QROL, lack of certainty lowers the grade. As opposed to the QCM where uncertainty is not much of an issue due to the lack of options.

- QCL incidence Practical : the incidence score is somewhat better. Apart from a couple exceptions most of the real grades have a small variance. This helps explain the very low 75 percentile of about 3 and the big number of erroneous cases above 6. Since this part of the exam is similar to a QCM but with only 3 options, it's relatively easy to answer the questions correctly. This explains the higher grades for most students.

When taking practical exams the students have images to look at. Certain aspects of the images need to be identified before answering the questions. Unfortunately with the data fetched from Cytomine, it's hard to determine whether or not the students were able to identify certain concepts when using Cytomine. But attempting to identify the features that are close to determining whether or not a student understood the course is a start. (Figures 4.13, 4.14, 4.15,4.16)

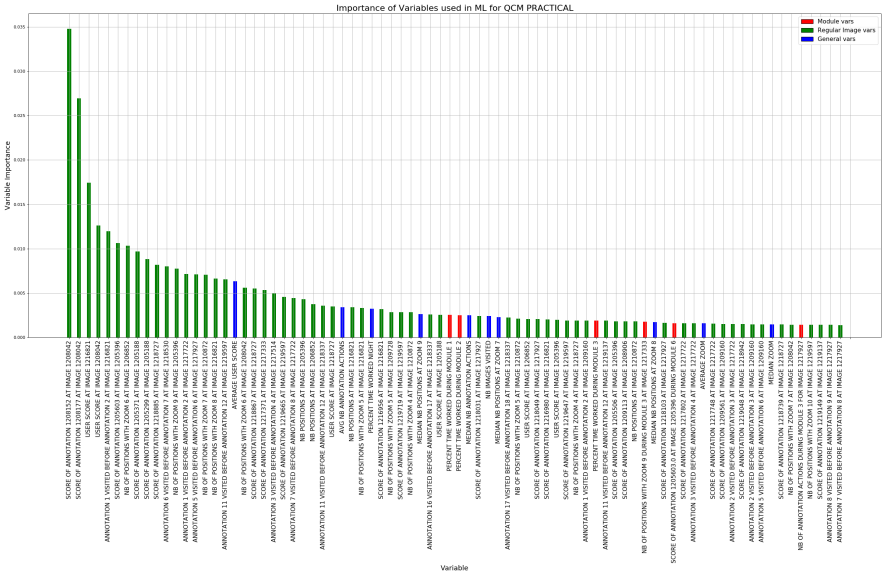


Figure 4.13: Feature Importance for the QCM Practical

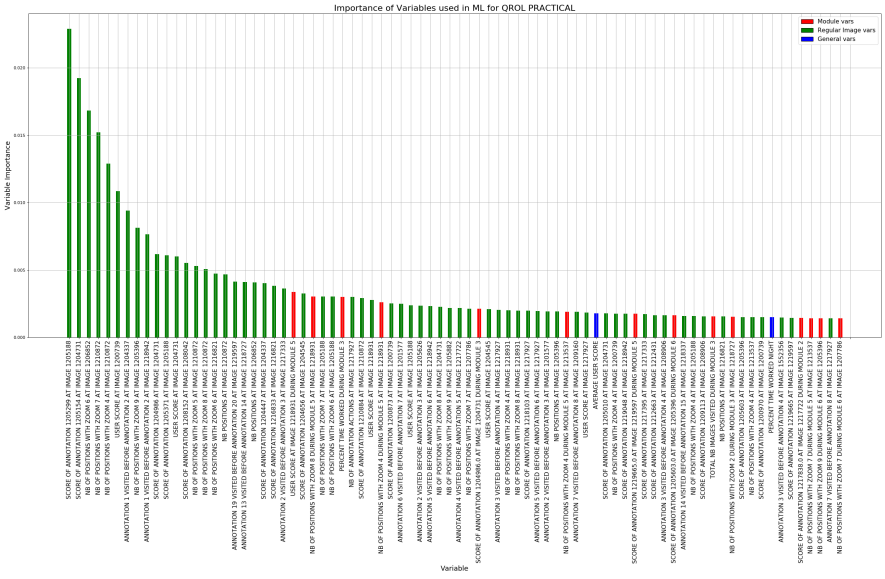


Figure 4.14: Feature Importance for the QROL Practical

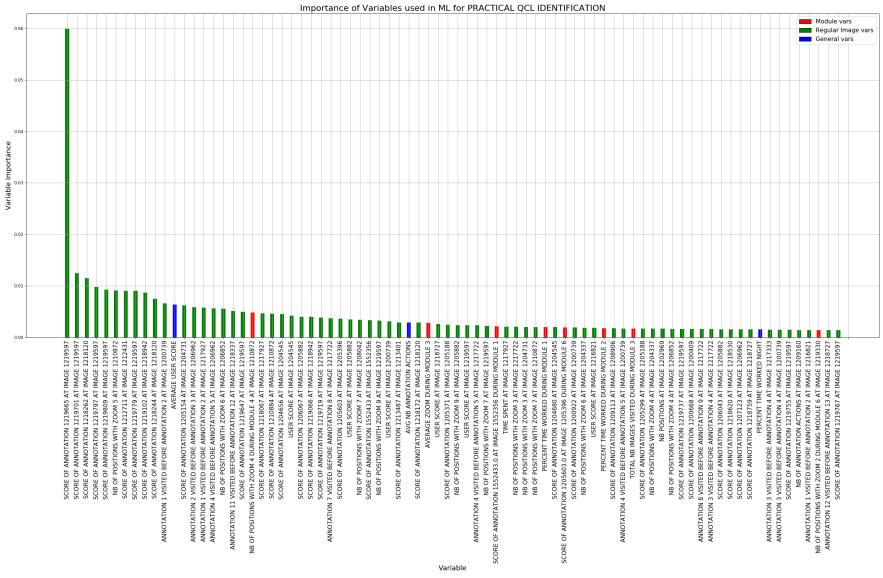


Figure 4.15: Feature Importance for the QCL Identification

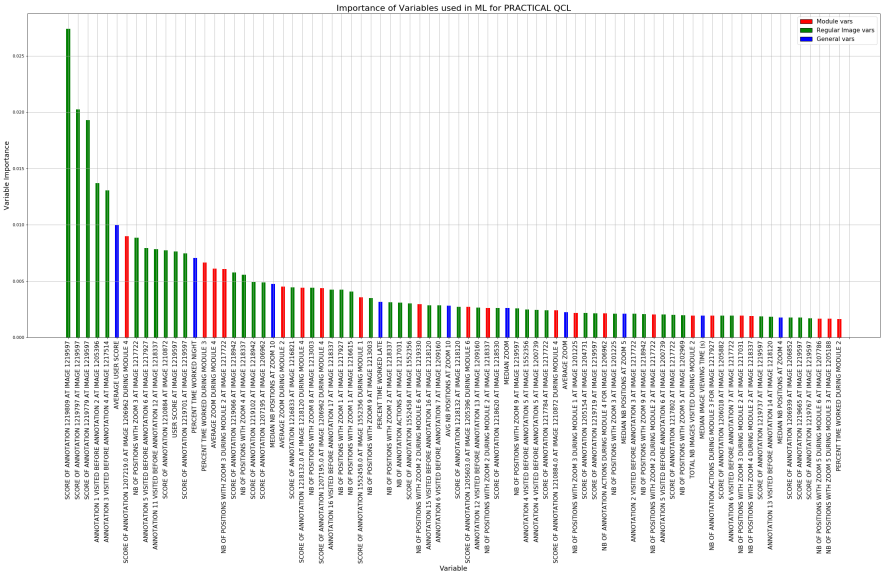


Figure 4.16: Feature Importance for the QCL Incidence

The results are different from the trial tests. For one, the Module features are relatively less impactful than before. This shows that the total work done throughout the semester is more important than working more during specific modules. It also is shown that the students spent the most time on Cytomine the day before the exam.Similar to the white tests models, the Average Score feature always shows up as a important feature. This variable seems to be the closest variable into determining whether or not a student understands the

course but it's still not perfect. In the end, some images and some modules have more impact than others. (Figures 4.17, 4.18, 4.19,4.20)

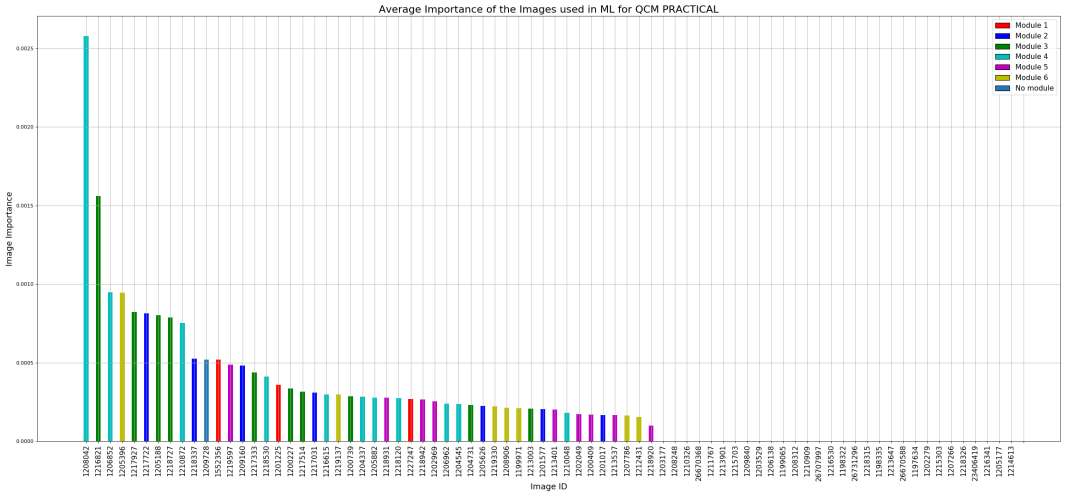


Figure 4.17: Image Importance for the QCM Practical

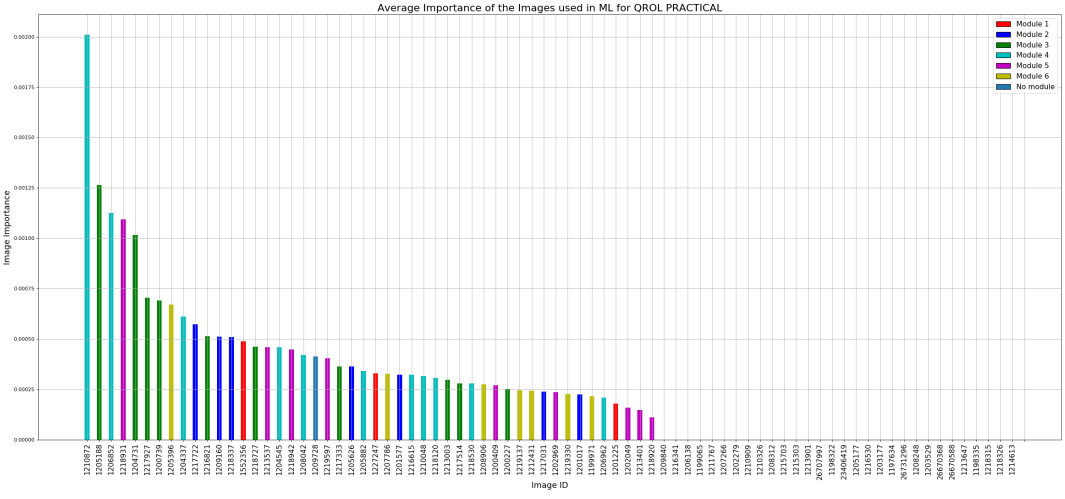


Figure 4.18: Image Importance for the QROL Practical

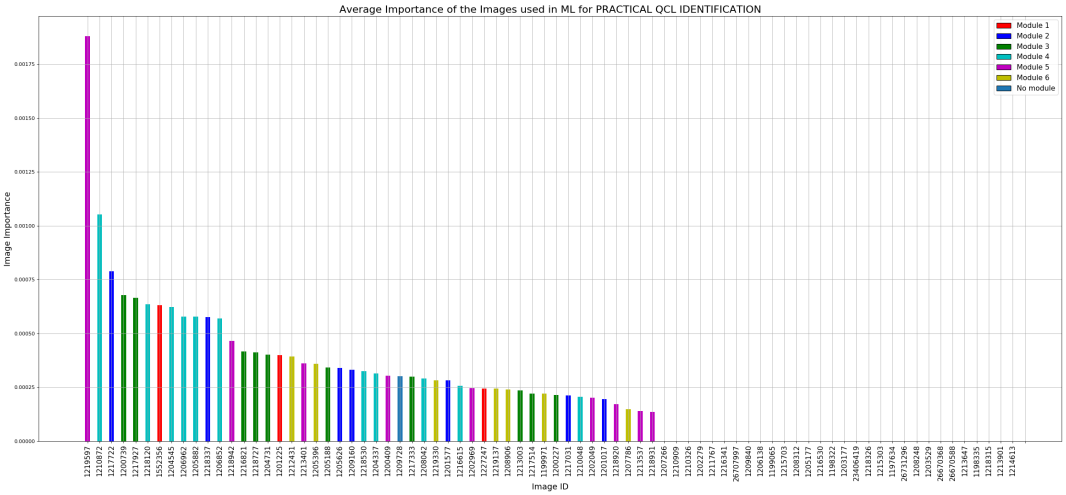


Figure 4.19: Image Importance for the QCL Identification

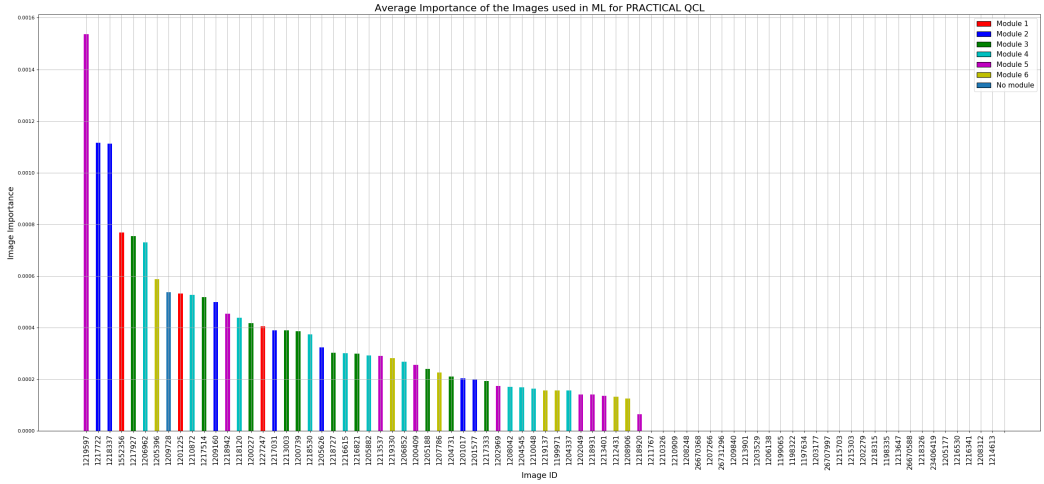


Figure 4.20: Image Importance for the QCL Incidence

It's hard to determine which modules were are more impactful as a whole.It seems that each module has at least one image that has a big impact on the result for each test. Like before, most these images are the ones dedicated to exercises. This supports the hypothesis that doing the exercises has a positive impact on the grades. But for example the image 1218120 does not appear at the top anymore. This shows that even tough the models follow the same patterns, it's still varies based on the contents of the exam.

Sometimes it's interesting to look for correlations between the expected result and some features. This helps to see its direct impact in estimating a grade. In fact, with most of the the features, a higher value should mean a better score. To test this, the Pearson Correlation was calculated for the top 6 features of the Practical QCM. (Figure 4.21)

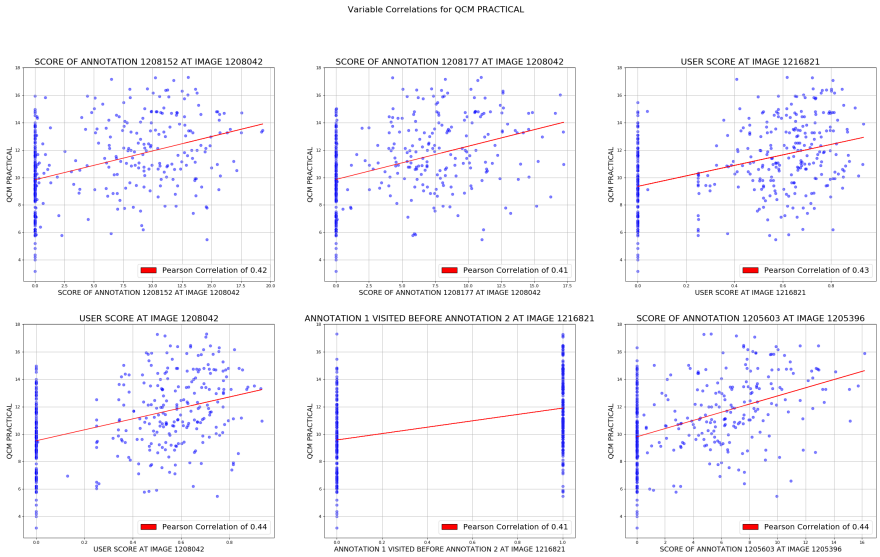


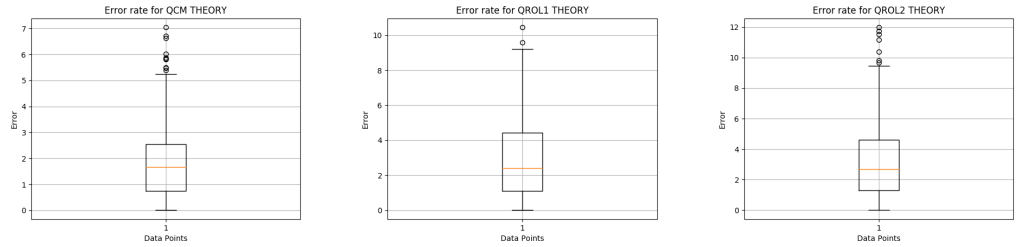
Figure 4.21: Pearson Correlation

Most of these features are score features which in short tells the model how well the user viewed the image in regards to the annotations. Therefore a higher value means a better observation of all the annotations. This correlation means that the more time the student is focused on annotations, the more likely that student is to get a better grade. This could explain the somewhat good Pearson Correlation for these variables.

After studying the results for the practical exams, a question that can be asked is whether or not this analysis can work on Theoretical Exam grades.

4.3.3 Theoretical Exam Grades

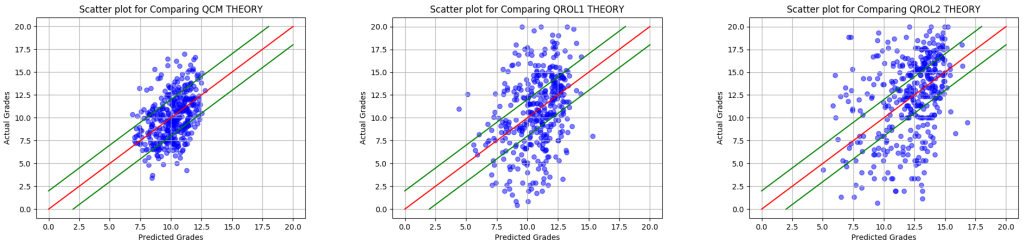
The same tests were launched for the theoretical grade results. The main difference for the theory portion is that there are no QCL tests and two QROL tests. The learning done on the QROL test results of the practical showed the worst results. Meanwhile the QCM provided the best results. This trend is likely to continue. (Figure 4.22)



4.22.1: Boxplots of the error values

	Score	Median Score	Score Difference	Real Median Grade	Median Est. Grade
QCM Theory	1.86	2.05	0.19	10.03	10.24
QROL1 Theory	3.00	3.30	0.30	10.87	11.05
QROL2 Theory	3.25	3.60	0.36	13.0	12.67

4.22.2: Discrete Results



4.22.3: Actual grades compared to Predicated grades

Figure 4.22: Results from cross-validation of the practical tests

As predicted, the QCM yielded the best scores while the QROL provided sub par scores. The assumptions made based off of the practical exams were correct. The Extra Trees has a easier time predicting QCM exams. But in this case, the QCM grades has a lower variance then the QROL.

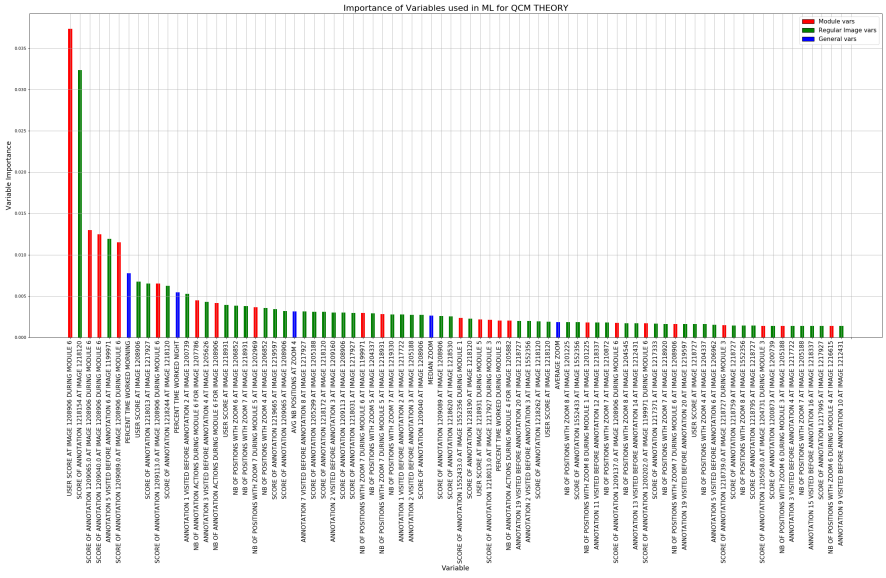


Figure 4.23: Feature Importance for the QCM Theory

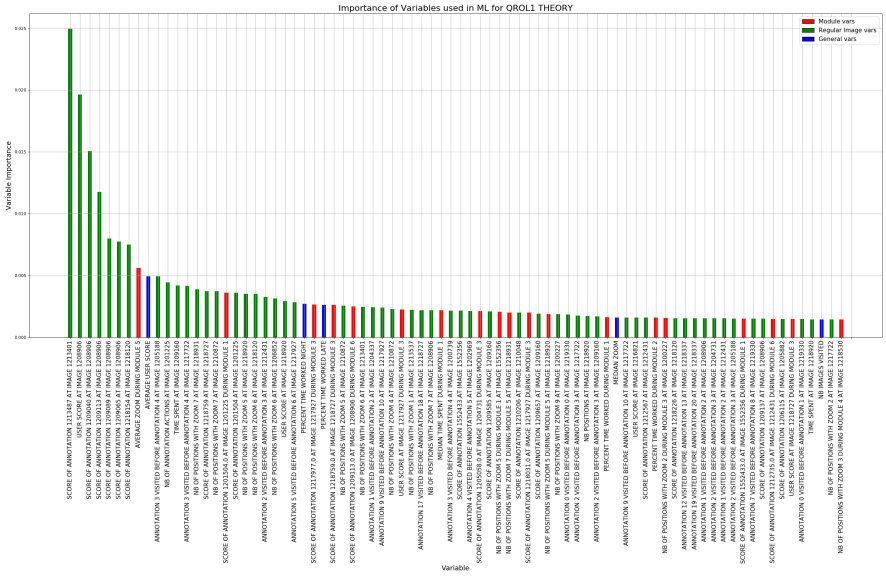


Figure 4.24: Feature Importance for the QROL1 Theory

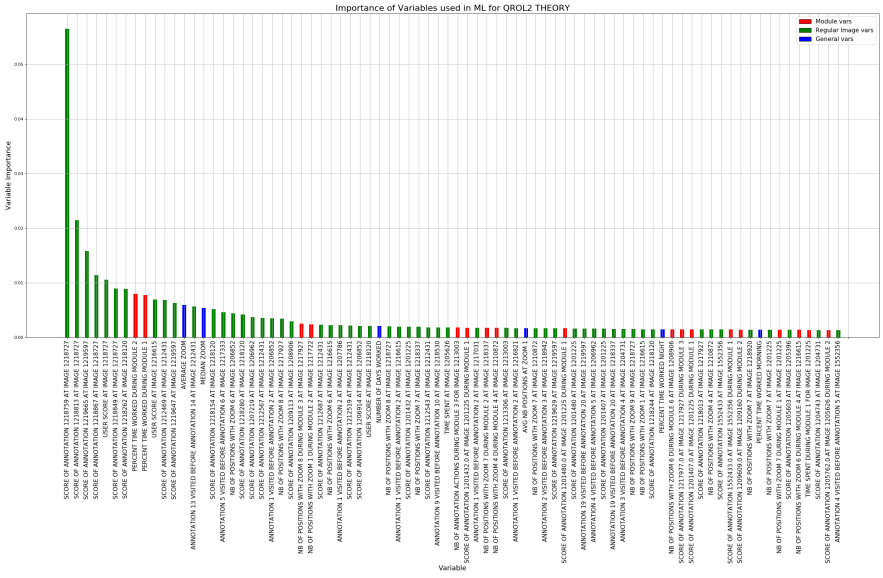


Figure 4.25: Feature Importance for the QROL2 Theory

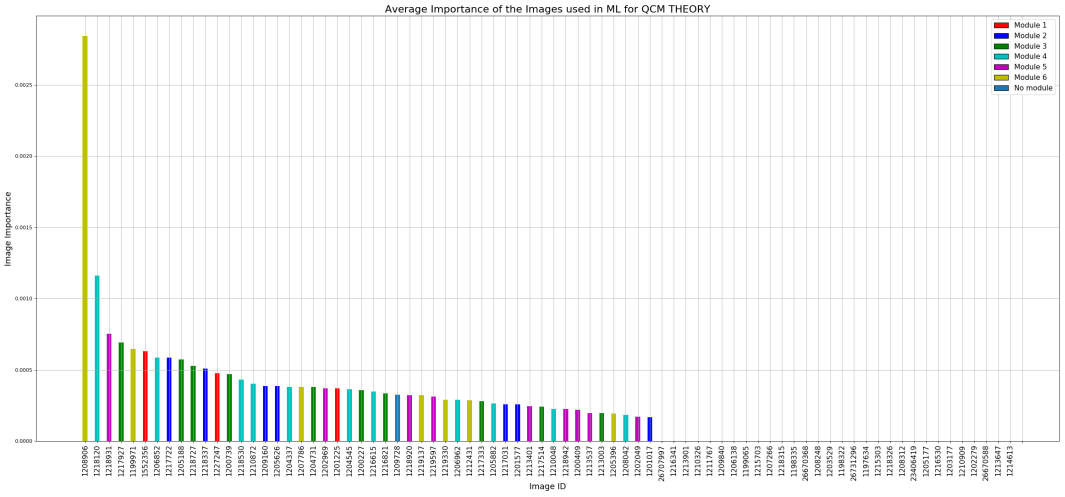


Figure 4.26: Image Importance for the QCM Theory

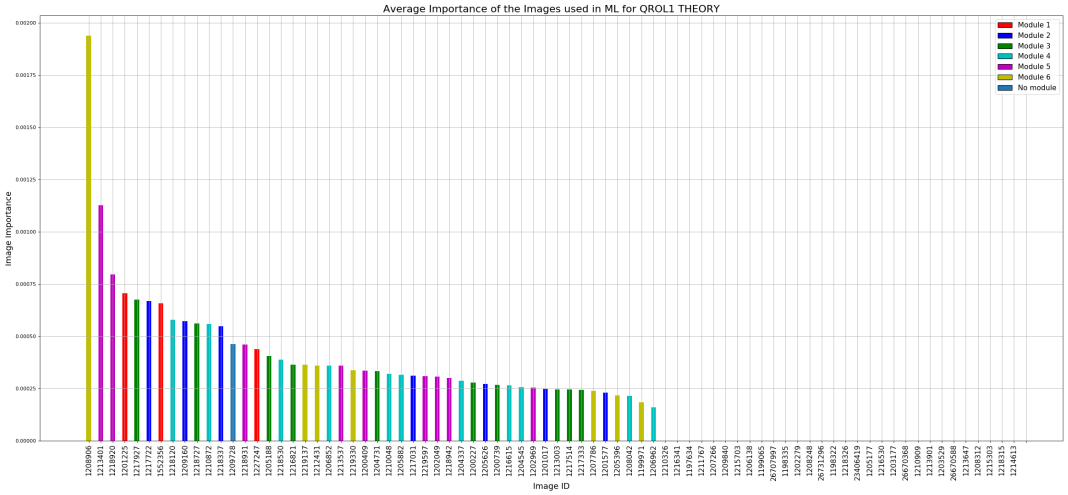


Figure 4.27: Image Importance for the QROL1 Theory

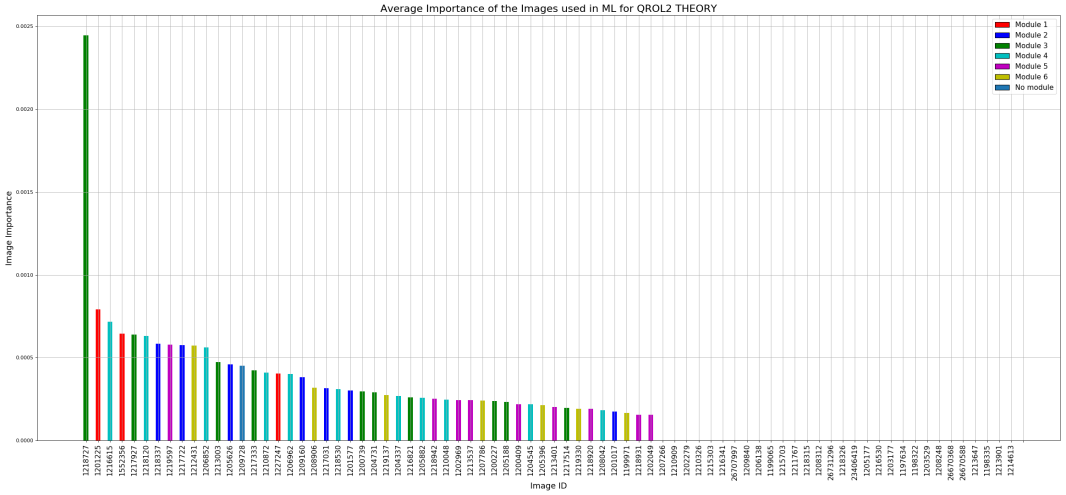
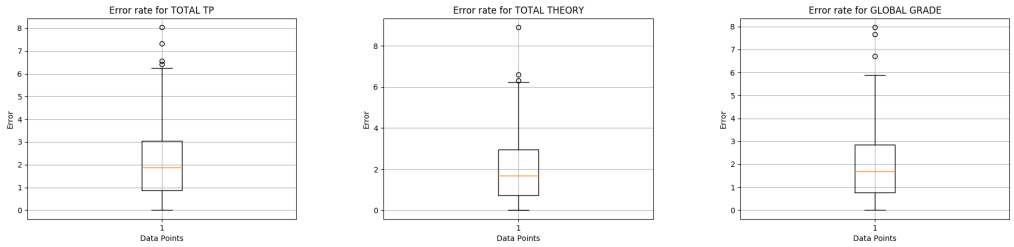


Figure 4.28: Image Importance for the QROL2 Theory

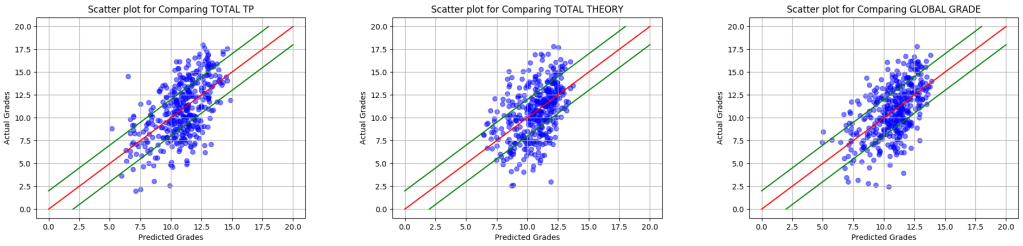
4.3.4 Global Grades



4.29.1: Boxplots of the error values

	Score	Median Score	Score Difference	Real Median Grade	Median Est. Grade
Total Practical	2.12	2.59	0.47	10.80	11.06
Total Theory	2.03	2.34	0.31	10.79	10.86
Global Grade	1.95	2.37	0.42	10.72	10.95

4.29.2: Discrete Results



4.29.3: Actual grades compared to Predicated grades

Figure 4.29: Results from cross-validation of the exam results

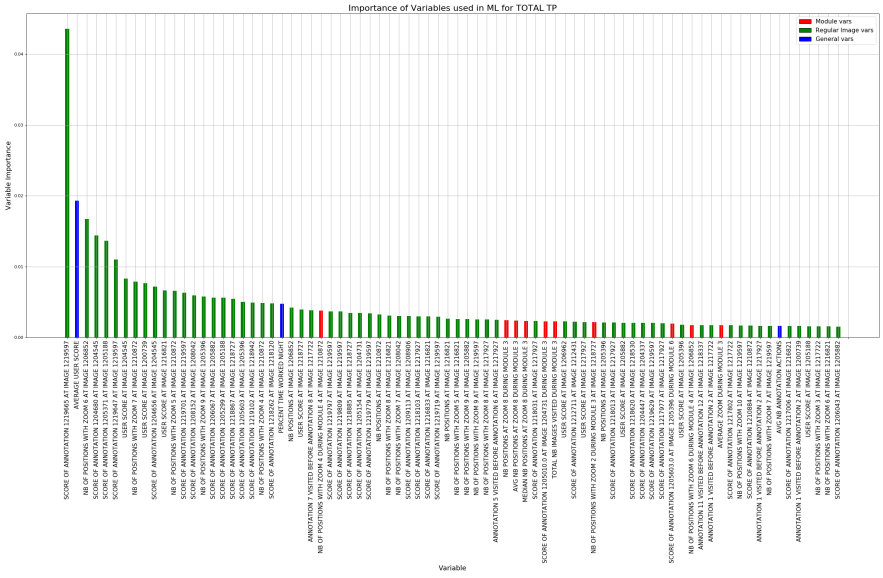


Figure 4.30: Feature Importance for the Practical

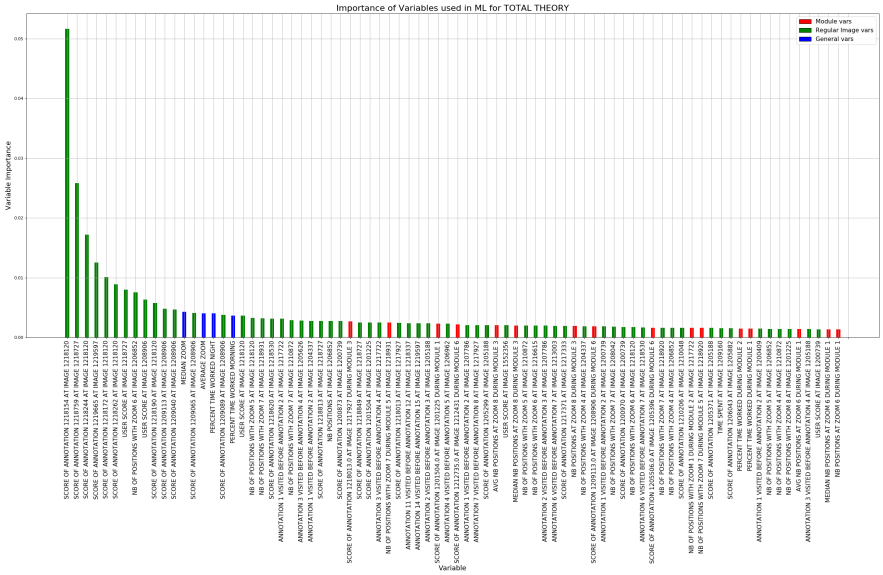


Figure 4.31: Feature Importance for the Theory

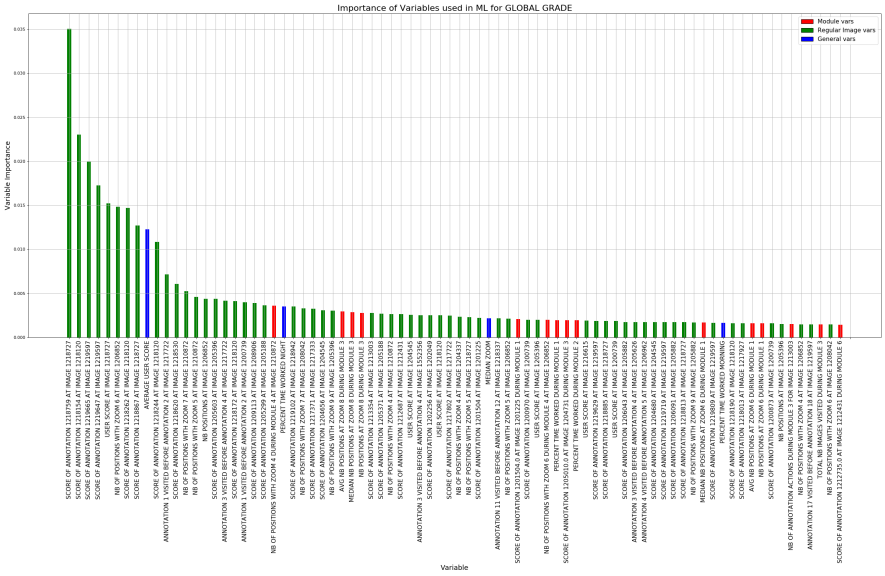


Figure 4.32: Feature Importance for the Total

4.3.5 Learning with additional Information

5 Discussion

6 Conclusion