Project Report

By Gautam Swaminath Ganesh

Abstract:

The aim of this study is to determine a method for understanding ensemble models through visual analytics. The method is used to identify and quantify the confidence levels between the models that constitute the ensemble. This is done using an interactive dashboard with visualizations to check the agreement and disagreement of the model's prediction in comparison with the true class. Evaluation of the project is performed with visualization as the core aspect of the user study. Scenarios are chosen with appropriate questions and goals and evaluated based on the results.

1.1 Introduction:

A general meta approach to machine learning seeks better productivity and performance by combining the predictions from multiple models. Ensembles, over the course of time, proved to be very effective over the use of single models and were evaluated using performance and robustness.

Although Ensembles proved to be a better option there still existed downsides or flaws, such as increased computational costs and complexity. And while using such complex models it came to light that there still existed certain exceptions such as mislabeling instances due to conflicts between models in the classification algorithm.

Through further research [1], it was found that the models that individually compile to form the ensembles have discrepancies between themselves. These discrepancies are also found to be the reason why such errors persist in the model. Thus the project aims to shed light on the discrepancy found between the constituent models and how trustworthy the models are.

In order to better understand the problem and have an in-depth view of the instances and determine the faults in the model, visualizations have been found to be a prime solution in such cases [2]. However, visualization of ensembles is difficult and are very sparse in number and provides a different purpose than what is expected for the current problem such as shape alignment, and is dedicated to multi-step data ensembles.

Thus in this project, we will be considering classification models such as Naive Bayes and KNN as the core models that constitute the ensemble and perform the classification on the iris dataset. The iris dataset is a standard dataset found on Kaggle [3] to determine the class of a flower through various parameters. This project is aimed to be completely dataset independent. Thus just for the sake of performing the analysis, the iris dataset is chosen.

Thus in order to visualize ensembles, they are considered to be multi-dimensional datasets, where each member represents one dimension ie. each member represents one or more data attributes. Thus using the existing methods of visualization of multi-dimensional data, visualization is created in order to fulfill the motive of representing the ensemble using confidence and disagreement as core factors.

A common approach to multidimensional visualization is to represent data elements with geometric shapes or glyphs. The visual properties of a glyph are varied to visualize its data element's attribute values [4]. Thus, a visualization based around scatter plots and glyphs may probably be a suitable alternative for this scenario.

1.2 Definitions:

These are some concepts that are recurring and play an important part in the project.

- *Ensemble:* In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

- *Confidence:* A machine learning probability score that tells you how confident the underlying algorithm is that it has extracted the correct value. The trouble with the confidence score is that if it is not 100%, you cannot reliably decide whether you need to look at the extracted data or not.

- *Disagreement:* this is a value determined by how much the models differ from one another with respect to classification accuracy.

- *Naive Bayes:* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

- *KNN:* K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new data points which means that the new data point will be assigned a value based on how closely it matches the points in the training set.

- *Glyphs:* In the context of data visualization, a glyph is any marker, such as an arrow or similar marking, used to specify part of the visualization. This is a representation to visualize data where the data set is presented as a collection of visual objects. These visual objects are collectively called a Glyph. The glyph is also the visual representation of a piece of data where the attributes of a graphical entity are dictated by one or more attributes of a data record.

1.3 Objective and Motivation

This paper will explain the uses of ensembles and the interactions between different classification models. It also shows the importance of glyphs and visualization techniques to further bolster the user experience. The expected reason for the user to employ the current application is to understand and utilize the information gained on the model which constitutes an ensemble. The model primarily focuses on the classification aspect of the project. This would imply that the user has the capability to understand classification models and also the relationship between multiple models. It also provides a brief understanding of the correlation between confidence and prediction accuracy. The result would be presented through a visualization. This would assist the users in better understanding the data and also offer customization capabilities and interactivity.

2.0 Related Work

The current application draws inspiration from several research fields. In this section, we will review and summarize only the most related work.

2.1 Ensembles

Ensembles are widely used models for increasing performance and have a proven record of being better than individual machine-learning models [5]. Ensemble models proved to be very effective over the use of single models and it was seen as such when multiple evaluations were conducted between them. There are two main reasons to use an ensemble over a single model, and they are related [5]; they are:

- *Performance*: An ensemble can make better predictions and achieve better performance than any single contributing model.

- *Robustness*: An ensemble reduces the spread or dispersion of the predictions and model performance.

Ensembles are used to achieve better predictive performance on a predictive modeling problem than a single predictive model. The way this is achieved can be understood as the model reducing the variance component of the prediction error by adding bias [4].

There is another important and less discussed benefit of ensemble methods is improved robustness or reliability in the average performance of a model. These are both important concerns in a machine learning project and sometimes we may prefer one or both properties from a model [4].

Research into ensembles has been extensive in the aspects of increasing the accuracy of the model and its robustness [4]. Papers that delve into the reasons for the errors are far fewer in number and are less explored in comparison [4]. Thus the project aims in providing a clear view of such errors through visualization techniques.

2.2 Ensembles to Model Comparison

To explore further reasons as to why the ensemble behaves in such a way a comparison of the constituent models are taken into account. This is clearly seen as a case of comparisons between multiple models. Model comparison means methods in which multiple machine-learning models are compared and evaluated in order to achieve a certain purpose or objective.

Research regarding multiple model comparison has been explored in depth with regard to classification models such as Naive Bayes and KNN in the fields of medicine [6/5] and pharmaceutical companies [7]. The research and analysis provided by these are given in-depth explanations with regard to model comparison and the ideal methodologies to compare multiple models [8].

An example regarding the comparisons of random forests is explored [9], through the tool known as Random Forest Similarity Map(RFMap), an interactive visualization tool that supports the explanation of RF models globally and locally by preserving the forest context. RFMap uses dimensionality-reduction techniques to visualize the forest of decision paths and explain the structural relationships between the data instances from the model's perspective. It also helps users visualize RF models with thousands of logic rules, one of the most scalable visualization solutions for RF analysis. The research papers provide a detailed guide and inform the readers of the different model comparisons and their methodologies

2.3 Importance of confidence and disagreement in model comparison

Model comparison methodologies generally involve the use of accuracy. But as we delve into a more complex and thorough understanding of the model factors such as confidence and disagreement come into the picture. Research regarding the behavioral aspects of humans with regard to their interaction with Artificial intelligence and Machine Learning models is well-studied and documented [1]. These research studies with the addition of studies that further stress the importance of confidence and disagreement help to explain the models and instances in a clear fashion.

There are also research studies that elaborate on the importance and methods by which confidence can be studied and used in machine learning models [10]. The results suggest that, in general, the confidence an ML model ties to its predictions significantly influences how many laypeople claim to believe in the model's individual predictions, but performance measurements like the model's accuracy on a held-out set of data and observed accuracy in practice have a greater influence on how often laypeople will actually follow the model as well as their self-reported overall trust in the model [1,10].

The exploration suggests a new method or in-depth analysis of the causes of such trust and an emphasis on the accuracy and confidence relationship [11]. These methods help us understand how confidence could be used and its relation with the user. Using this a framework regarding prediction accuracy and confidence is experimented on and is seen to be directly correlated.

2.4 Visualization

Although the previous section could explain and evaluate the confidence and disagreement between models, the goal of the analysis is still an aspect that is complex and requires more detail in the model. Thus visualization is utilized.

Visualization is used to provide an in-depth view of each instance of the dataset and determine the fault. However, visualization of ensembles is difficult. There have been few studies that have taken this approach and provided a different purpose than what is expected for the current problem [2].

Existing visualization techniques could conceivably be used to visualize ensembles. Since our ensembles contain volumetric data, some extension of direct volume rendering might be appropriate. An ensemble could be viewed as a multidimensional dataset, with each member representing one dimension [4].

Another approach is to treat an ensemble as a multidimensional dataset, where each member represents one or more data attributes. Existing visualization techniques could then be used to visualize the ensemble as a multidimensional image [4].

However, the method or type of visualization that would be appropriate was not yet known and was a hurdle. The research regarding the types of ensemble visualization provided key insights regarding the type and method of the visualization to be used [5]. The study proposes two new methods for visualizing ensembles that contain numerous member datasets: a pairwise sequential animation technique that combines subsets of members using visibility control, and a screen door tinting technique that presents differences between members using screen space subdivision and saturation tinting. Using this, a sequential technique is opted in the current study for visualization and in the form of scatter plots. This form of visualization will play a significant role in providing a form of visualization that prioritizes values such as confidence and disagreement over the conventional use of accuracy. This form of visualization is also easily scalable because it is possible to segregate and view a part of the instances at a time.

2.5 Glyphs and their role

A common approach to multidimensional visualization is to represent data elements with geometric shapes or glyphs. The visual properties of a glyph are varied to visualize its data element's attribute values. Many multidimensional visualization techniques use glyphs.

Early examples include Chernoff faces [5], where faces represent data elements and a face's expression represents attribute values or starfields, where rectangular glyphs are arrayed in 2D to visualize multidimensional data in a scatterplot-like fashion. More recent approaches apply perceptual guidelines [5] on the use of color and texture to general highly salient glyph patterns.

In order to make sense of high-dimensional data, data scientists need to enter into a dialogue with domain experts this was the Glyphboard [12], a visualization tool that aims to support this

dialogue. Glyphboard is a zoomable user interface that combines well-known methods such as dimensionality reduction and glyph-based visualizations in a novel, seamless, and integrated tool. While the dimensionality reduction affords a quick overview of the data, glyph-based visualizations are able to show the most relevant dimensions in the data set at one glance.

Thus upon experimentation using the information obtained using this research [5,12], there were two glyphs that performed ideally and were apt for the task. The first one was the flower glyph. A flower glyph is one in which the attributes are represented using the sizes and colors of the multi-faced petals. The second one is the polygonal glyph, which is used in a similar fashion as the flower glyph, through the length of each side and the color of its shape.

Using this, an interface is created to show and customize the visualization capabilities of glyphs in a user-friendly manner. Flower glyphs were also chosen as they were seen to be the best at describing the conditions in a clear and user-friendly manner. These glyphs are symbolically used to represent the results of each instance once they have been classified using the various ML models. The results obtained from the machine learning models are then used in the visualization using these selected glyphs. The predicted accuracy and the parameters associated with it are taken into account when calculating the confidence of the model. This confidence value is then added to the instance through the use of either flower glyphs or polygonal glyphs.

2.6 Dimensionality Reduction and DGrid

Dimensionality reduction is a machine learning (ML) or statistical technique for reducing the number of random variables in a problem by obtaining a set of principal variables. This process can be carried out using a number of methods that simplify the modeling of complex problems, eliminate redundancy and reduce the possibility of the model overfitting and thereby including results that do not belong [13].

The process of dimensionality reduction is divided into two components, feature selection, and feature extraction. In feature selection, smaller subsets of features are chosen from a set of many-dimensional data to represent the model by filtering, wrapping, or embedding [10]. Feature extraction reduces the number of dimensions in a dataset in order to model variables and perform component analysis.

There are two different types of dimensionality reduction techniques that have been used in this project. The first one is PCA dimensionality reduction. Principal Component Analysis (PCA) is one of the most commonly used unsupervised machine learning algorithms across a variety of applications: exploratory data analysis, dimensionality reduction, information compression, data de-noising, and plenty more [13].

The second one is t-SNE dimensionality reduction, t-SNE is mostly used to understand high-dimensional data and project it into low-dimensional space (like 2D or 3D). That makes it extremely useful when dealing with CNN networks. One of the major differences between PCA and t-SNE is it preserves only local similarities whereas PCA preserves large pairwise distances

to maximize variance. It takes a set of points in high-dimensional data and converts it into low-dimensional data [13].
To further assist the visualization overlap removal of Dimensionality Reduction, Scatterplot Layouts [13] were used. Dimensionality Reduction (DR) scatterplot layouts have become a ubiquitous visualization tool for analyzing multidimensional data items with presence in different areas. Despite its popularity, scatterplots suffer from occlusion, especially when markers convey information, making it troublesome for users to estimate items' groups' sizes and, more importantly, potentially obfuscating critical items for the analysis under execution. Despite the good results of post-processing techniques [14], the best methods typically expand or distort the scatterplot area, thus reducing markers' size (sometimes) to unreadable dimensions, defeating the purpose of removing overlaps. Using this research paper the method to remove DR layouts' overlaps faithfully preserves the original layout's characteristics and markers' sizes.

3.0 Methodology

3.1 Preprocessing

The data that is taken into consideration is multidimensional data and thus requires preprocessing. Therefore we perform a standard routine check of the data to see if there are any missing values or errors on the dataset. They were followed by a dimensionality reduction which is paramount for a multidimensional dataset. A PCA dimensionality reduction is performed by default with the option for t-SNE. The data is then converted into a machine-readable format and is utilized by machine learning models.

3.2 Machine Learning Models

As per the definition of ensembles, it consists of multiple machine learning models. The visualization aims to classify and understand the individual models before the voting committee stage. Thus for the sake of this project, there are two models namely the Naive Bayes model and the KNN model. These models were chosen as they are the most commonly known classification models that have a consistent and trustworthy output.

The models are then utilized in order to obtain the predicted class and the true class. Using this the prediction accuracy, confidence and disagreement are then calculated. These values are then scaled from zero to one for better understanding.

3.3 Creation of Scatter plots

The Visualization utilizes a dashboard. The screen is divided into multiple SVGs, they are Scalable Vector Graphics is an XML-based vector image format for defining two-dimensional graphics, having support for interactivity and animation. In the SVG a scatter plot is displayed. The scatter plot is then further color coded to the predicted class.

3.4 Creation of Glyphs

Before the creation of glyphs, individual hover circles are created around each instance of scatter plots. These hover circles act as boundaries and enable safe and easy access to each instance in the graph. Using these hover circles glyphs are created with the circles acting like boundaries. This method is repeated for each and every instance in the considered dataset. The hover circles' diameter is set to one unit in length and 0.1 unit in thickness. Thus being able to act as a thin outline for the glyphs that will be created within the circle.

The color of the glyph is based on the predicted class and the outline of the glyph is based on the true class. the outline is set to 0.1 unit and belongs to the same area of the glyph. The size is set in such a way that it is only visible but does not affect the size of the petals or the polygon. In the case of the flower glyphs, the number of petals represents the different parameters that are determined after the dimensionality reduction. The length of each petal is affected by the disagreement between models. Ie. The larger the disagreement smaller the length of each petal and vice versa.

In the case of polygonal glyphs similar to the flower glyph, the number of sides is determined by the different parameters found after the dimensionality reduction, and the length of each side is determined by the disagreement between models. This is useful in order to see the imbalance or the error in the models.

3.5 Iteration 1

Using this algorithm we are then able to overlay the flower glyphs on top of the scatter plots. This will fulfill the objective (To visualize ensembles utilizing confidence and disagreement as core parameters) as the main focus of the project. Through the view present, we are able to identify the incorrectly labeled instances and will be able to see which model is more trustworthy and by how much using the disagreement factor. This can be identified using the height of the petals, The longer the petal, the greater the confidence between each model. This visualization can be used to determine the classes for each instance. Each instance is labeled and can be uniquely identified in the search bar utilizing the id number of the instance. This can be seen in Fig.1.

Fig.1 represents the scatter plot, overlaid with the flower glyphs.

3.6 Challenges Faced

The current iteration faced issues of importance and was not able to convey the message in a clear manner.

- *Faults 1.1*: The current model experiences certain issues. To start with the total number of instances that can be visualized for user testing is 700 instances. Thus it would not be feasible to visualize large datasets for eg. the MNIST dataset which contains over 70,000 instances.

- *Solution 1.1*: To combat this issue a K Fold validation set is created by which a validation set the size of only 700 instances is taken into account. Using this we are able to visualize the results of the classification algorithm by dividing the dataset into smaller batches. Due to the high volume of data, the visualization will also take time to be generated. The data spacing will be limited and resizability of the instances will no longer be possible.

- *Faults 1.2*: Once the visualization is complete it can be seen that the scatter plots are overlapping with one another. With the addition of flower glyphs, this is even more apparent.

- *Solution 1.2*: To combat this issue an algorithm known as a DGrid is utilized to remove the overlapping by placing each instance into grids, where the distance between each instance in the grid is equal to the distance between instances without a DGrid. Using this we are able to remove any form of overlap and obtain a clear picture of what the classification looks like. This can be seen in Fig.2.

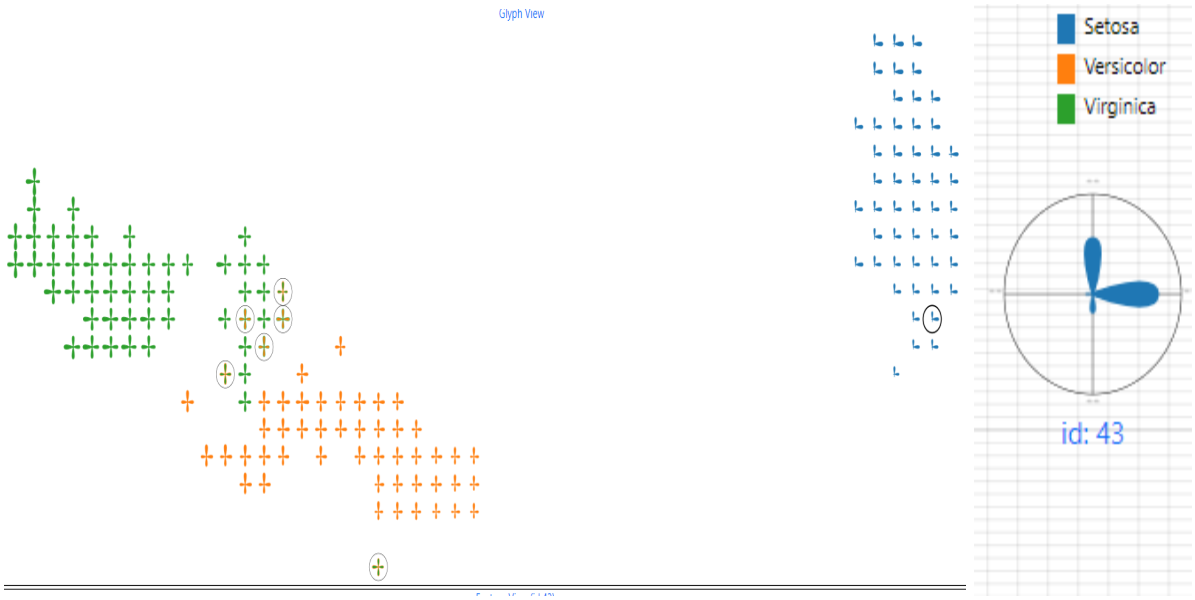

Fig.2(Left) Represents the scatter plots with the introduction of DGrid.
Fig.3(Right) represents the overlay created to visualize the instance better.

- *Faults 1.3*: When taking into account a large dataset, visualization is crowded and each instance although visible is not big enough to be properly analyzed. The instances are not visible and multiple clicks are required to select the instance that is required

- *Solution 1.3*: To combat this issue an overlay is created on the right top corner to display the glyph which is selected. Using this a clear view of the instance is visible and further analysis can be performed. This can be seen in Fig.3.

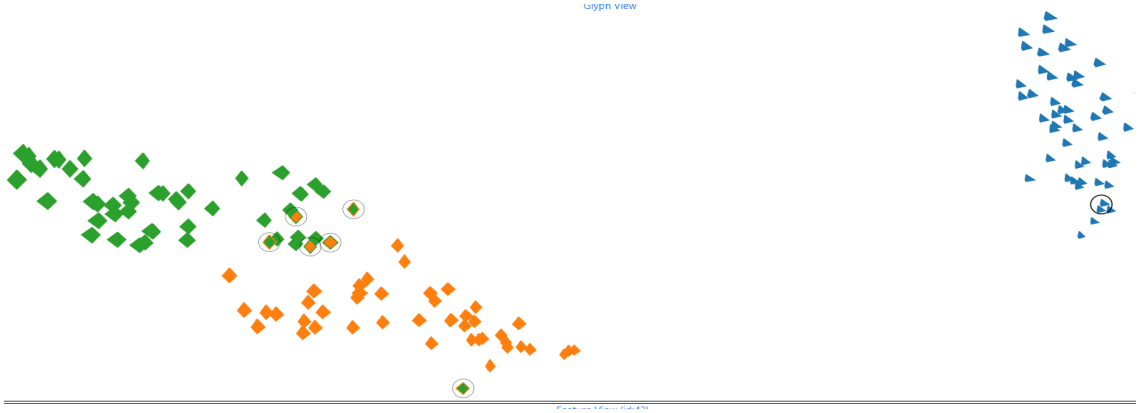

Fig.4 represents the scatter plots overlaid with polygonal glyph

- *Faults 1.4*: Utilization of the flower glyph could be considered complicated and hard to understand.

- *Solution 1.4*: A polygonal glyph is created for an alternate visualization. The rules that are used for the flower glyph are also used for the polygonal glyph. This can be seen in Fig.4, and Fig.5.

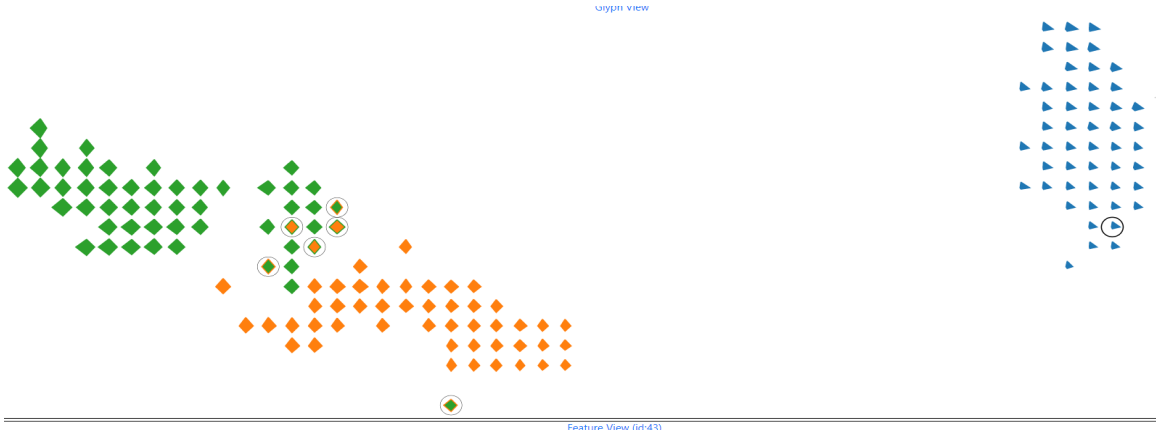

Fig.5 Represents the scatter plots with the introduction of DGrid.

3.7 Iteration 2

In this section, we focus on the QoL (Quality of Life) changes and the additional features added in order to provide a better viewing experience and to understand the dataset completely. These features are implemented in various forms such as a taskbar and a parallel coordinate chart. Solution 1.3 which provided an individual view into the instance is also made to appear on the same screen instead of a different screen. A tooltip is also created in order to optimize certain features such as minimizing, magnifying, and resetting the screen to the default size.

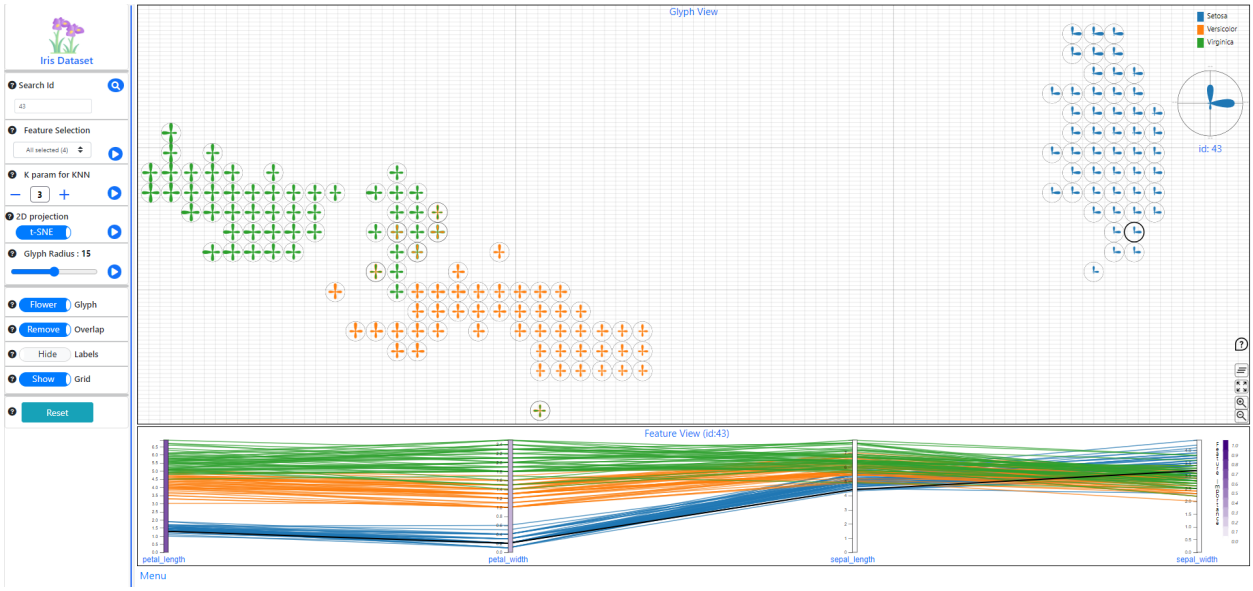

Fig.6 represents the overall visualization dashboard post changes.

3.8 Taskbar

The taskbar is created to better encapsulate the different quality-of-life changes that are made for customizability and a better viewing experience.

- *Search ID:* A search bar that is created in order to select an individual instance without the need to search for the instance manually.

- *Feature selection:* A drop-down search menu for the sole purpose of selecting relevant features for the current iteration of the model. By default, all the features are selected.

- *KNN variability:* By default, the K value for the KNN model is set to 3. This function provides an additional option to further customize the KNN model by providing a slider to increase or decrease the K value.

- *2D Projection:* A toggle button that is created in order to provide a multi-facet view of the different projections available when using different dimensionality options. By default the dimensionality reduction that is chosen is t-SNE. The other option provided is the PCA dimensionality reduction. By default, the 2D projection method is chosen to be t-SNE.

- *Glyph Radius:* A slider function created to increase the size of all glyphs but fit in the same window size. This magnification is done using the Dgrid algorithm to retain the physical qualities and increase the size of each icon. By default, the size is set to 15 units.

- *Overlap:* A toggle function was created to apply and remove the Dgrid functionality. This function is used if the user wishes to see the original view ie. the model without the overlaps between glyphs removed. By default, this function is set to enable overlap removal.

- *Labels:* A toggle function was created to add labels or remove them. The label contains the id number of the instance, this is done in order to further locate or search for a particular instance. The value is set to hide labels by default.

- *Show Grid:* A toggle option was created in order to further dissimilate and view the glyphs in a more clear manner.

- *Reset:* A button function to revert all changes to default.

- *Menu:* a button to hide or display the taskbar

This taskbar can be seen in Fig.6 located on the left-hand side. The menu button is found on the bottom left-hand side.

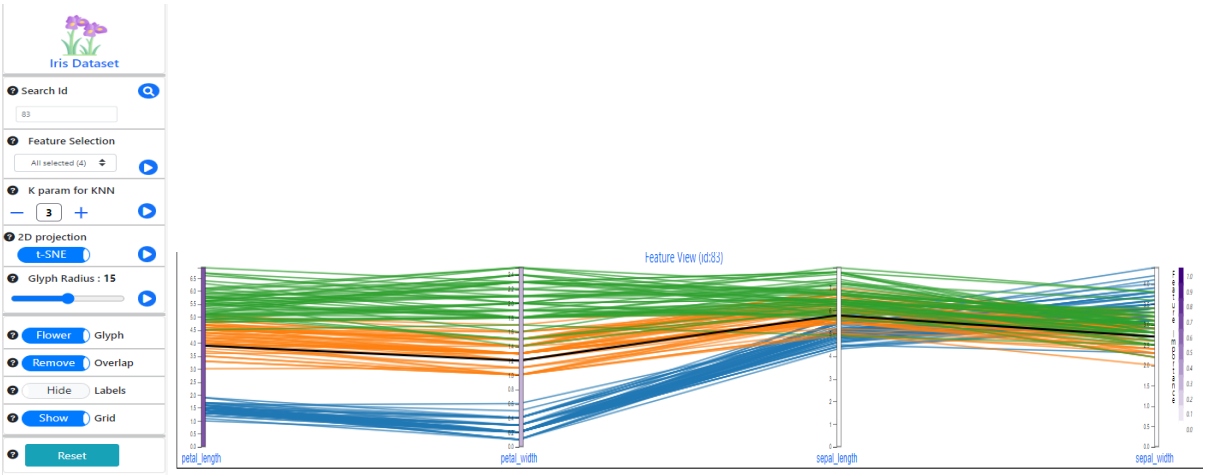

Fig.7 (Left) Represents the taskbar and the default selected options
Fig.8 (Right) Represents the Parallel coordinate chart and the selected parameter with Id. 83

3.9 Parallel Coordinate Chart

Through these forms of visualization, information regarding classification and confidence can be understood. This however does not do justice in explaining the dataset as a whole. Thus we require an alternate method through which the dataset can be explained. Thus a parallel coordinate chart is created. A parallel coordinate plot maps each row in the data table as a line or profile. Each attribute of a row is represented by a point on the line. This makes parallel coordinate plots similar in appearance to line charts, but how data is translated into a plot is substantially different.

The main advantage offered by parallel coordinates is the representation of high-dimensional data as a 2-dimensional visualization. As data is represented in the form of a line, it becomes easy to perceive the trend shown by data entries from the visualization.

The parallel coordinate chart is created with the features that constitute the model as its core. This can be seen in Fig.8. Using this information we can see the general trend of the dataset that would lead to a particular classification. The different classes are marked with the color scheme used to determine the glyphs. An additional feature is implemented where if one instance is selected on the glyph chart, its corresponding instance in the parallel coordinate chart is colored black and we can further inspect the characteristics of the instance.

3.10 Cluster analysis

To be done

4.0 Evaluation

Since this is a visualization-based project, evaluation cannot be done using features such as numerical accuracy. Thus we require an extensive user study-based evaluation. This requires a

questionnaire to be formed and the model is tested by others members of society that have primary knowledge of fields that are associated with this project.

4.1 Intended Goals

The primary objective of the user study is to identify faults in the current iteration of the model and perform an unbiased evaluation of the model. In an ideal case, the model is deemed to be satisfactory from a user standpoint. This would imply that no major flaws exist in the model's design and objectives. The evaluation will begin once a brief introduction to the prototype is given to the users, and a list of subjective and descriptive questions is provided.

The evaluation aims to validate the prototype and find issues with complex flows. It also serves the purpose of obtaining unbiased user opinions and getting their insights that can help create a better overall user experience.

The evaluation will be conducted in between-subject testing. Between-subjects is a type of experimental design in which the subjects of an experiment are assigned to different conditions, with each subject experiencing only one of the experimental conditions.

Compared to within-subjects studies, between-subjects studies have shorter testing sessions. Test participants who are assigned one website to test will be able to complete the usability test faster than those who need to test two (or more) websites, making the between-studies approach ideal for remote usability tests.

4.2 Method of Evaluation

User experience is getting more and more important for digital products nowadays. When not accounting for usability testing, your website may seem out of place, order, or even date. With usability testing studies you can ask your users directly about their opinions, and make decisions based on facts.

In this project, task-driven usability tests are preferred over the other evaluation methodologies. The basic requirement of all customers is to easily navigate to the product, and information they are looking for effortlessly. They expect an issue-free experience while using the website. However, this ideal case is not likely in the current situation.

Task-oriented usability studies aim to combat this problem. Such studies combine qualitative research methods to provide you with in-depth explanations, and as much context as possible while not forgetting about important quantitative metrics and statistics. They take into account that your users need to accomplish specific goals on your website and they should be able to do so easily.

4.3 Evaluation Environment

When conducting an evaluation based on a user study, the environment in which it is conducted is equally important. The environment in this case refers to the conditions in which the evaluation will take place. These parameters are the following.

- *Sample Size:* The sample size to be expected is 50-60 individuals that possess adequate knowledge of the subject matter. This is found to be the average number of users for an evaluation.

- *Dependant Variables:* The main features that could change based on the user study are the glyphs and the taskbar. Certain features may be removed or added based on user feedback.

4.4 Tasks

The evaluation will begin once a brief introduction to the prototype is given to the users, tasks that the user should perform are provided and a list of subjective and descriptive questions is provided.

- *Task 1:* The user is to analyze the data that is present and identify the nodes in which there is a discrepancy. This would include incorrect classification, disagreement between models, and if so, ranking the order in which the classes have been classified.

- *Task 2:* Check which method of representation is more clear for the user. The users are tested in alternate pairs being presented with the flower glyph visualization first and the others being tested with the polygonal glyph. This task would assist in identifying preferred configurations and better visualization methods.

- *Task 3:* Identify the particular instance where the disagreement is highest and the corresponding feature attributes.

4.5 Evaluation Testing

The core objective of this evaluation is to determine which version of the design is more suitable for the user's experience. Thus the A/B test was chosen as the most effective strategy as the current iteration of the project provides the user with two different visualizations based on the unique glyphs. The main advantage of A/B testing is that it allows the users to choose the design that provides the most amount of information with the least amount of effort and difficulty. It also provides a numerical analysis for the two different versions A and B.

The metrics used to determine a website's efficiency are:

- *Discrete metrics or binomial metrics*. Only 0 and 1 are the possible values.
  In this case, the user’s ability to successfully complete the task is tested to valid or not.

- *continuous metrics also called non-binomial metrics*, the metric may take continuous values that are not limited to a set of two discrete states.
  In this case, the time taken by the user in order to complete all the tasks are taken into consideration.

With the data we collected from the activity of users of our website, we can compare the efficacy of the two designs A and B. Simply comparing mean values wouldn't be very meaningful, as we would fail to assess the statistical significance of our observations. It is indeed fundamental to determine how likely it is that the observed discrepancy between the two samples originates from chance.

From the tasks that the users are designed to accomplish, two sets of data are obtained. Using this data we aim to calculate the statistical significance of each visualization. Using this, it is possible to determine how much one model is better than the other. In order to determine the observed discrepancy between the two different visualizations a two-sample test is conducted.

The initial hypothesis (H0 or null hypothesis) is that the two visualizations are equally effective to achieve the set goal. Using the two-sample test, the statistical significance is measured using the p-value given by the formula in Fig.9. This would indicate the probability of generating a discrepancy that is at least equal to the one collected by the tests conducted.

$$p_{val} = p(\text{data at least as extreme as actual observation} \mid H_0)$$

Fig.9 represents the p-value formula

Now, some care has to be applied to properly choose the alternative hypothesis Ha. This choice corresponds to the choice between one- and two-tailed tests.

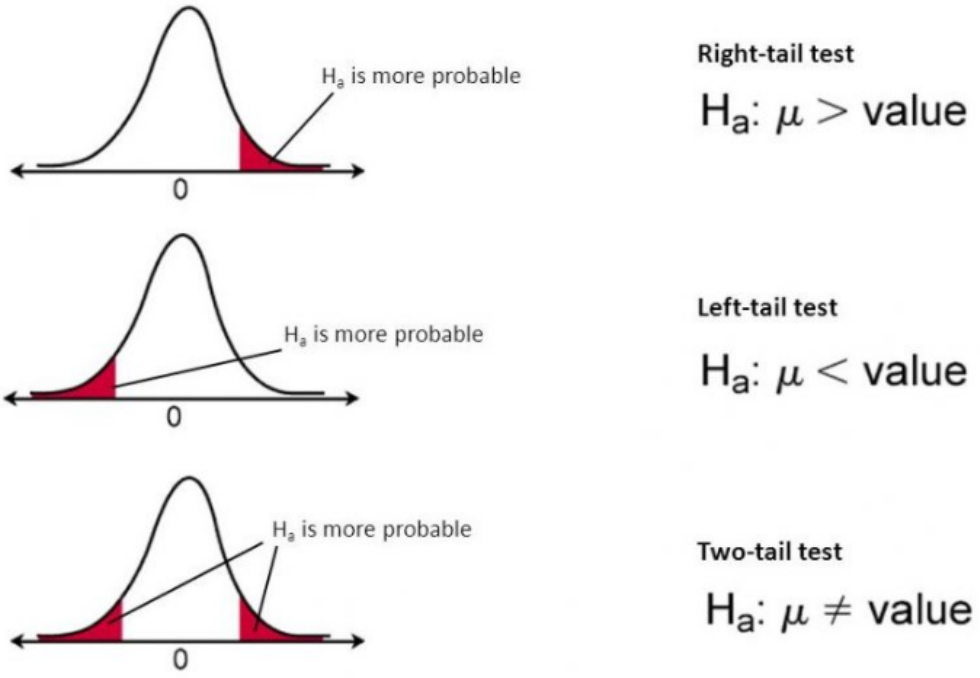

Fig.10 represents the one-tail test and two-tail test

From further introspection, it is found that the two-tailed test is much more preferable than the alternative because we do not definitively which visualization is better, model A or model B [13]. Therefore the alternative hypothesis (Ha) would consider that models A and B have different efficiency. The p-value is therefore computed as the area under the two tails of the probability density function p(x) of a chosen test statistic on all x' s.t. p(x') <= p(our observation). The computation of such a p-value clearly depends on the data distribution. So we will first see how to compute it for discrete metrics, and then for continuous metrics.

4.6 Discrete metrics and Fisher's exact test

The first parameter that is tested is the Discrete metrics, the test that is conducted is Fisher's exact test. In this case, the user's ability to perform the tasks successfully is assessed. Thus the data that is obtained will be in the following fashion. Let X be the number of observations in visualization model A and Y be the number of observations in visualization model B. This can be seen in Fig. 11. Using this data a 2x2 matrix ie. a contingency table is then obtained this can be seen in Fig. 12.

```
- Observations:
  - Version A: = [1 0 0 1 1 1 1 0 0 0 1 0 0 1 1]
  - Version B: = [0 0 0 0 0 1 0 0 0 0 1 0 0 0 0]
```

- Contingency table:

	A	B
click	7	13
no click	8	2

Fig.11 user task observations recorded (Left)
Fig.12 contingency table (Right)

Although using this matrix it is possible to inspect and deduce which model is better, it does not provide a statistical disparity between the two visualization models. Therefore using this contingency table we can use Fisher's exact test to compute the p-value and test the hypothesis [14].

The null hypothesis states that Visualization models A and B have the same efficiency, The probability of this event occurring can be calculated using the Hypergeometric distribution.

	A	B	tot
click	a	b	a+b
no click	c	d	c+d
tot	a+c	b+d	

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad p(M) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Fig.13 Hypergeometric distribution of possible outcomes

Using this formula in Fig.13 we can obtain the area under the two-tailed graph and determine the p-value as seen in Fig.14.

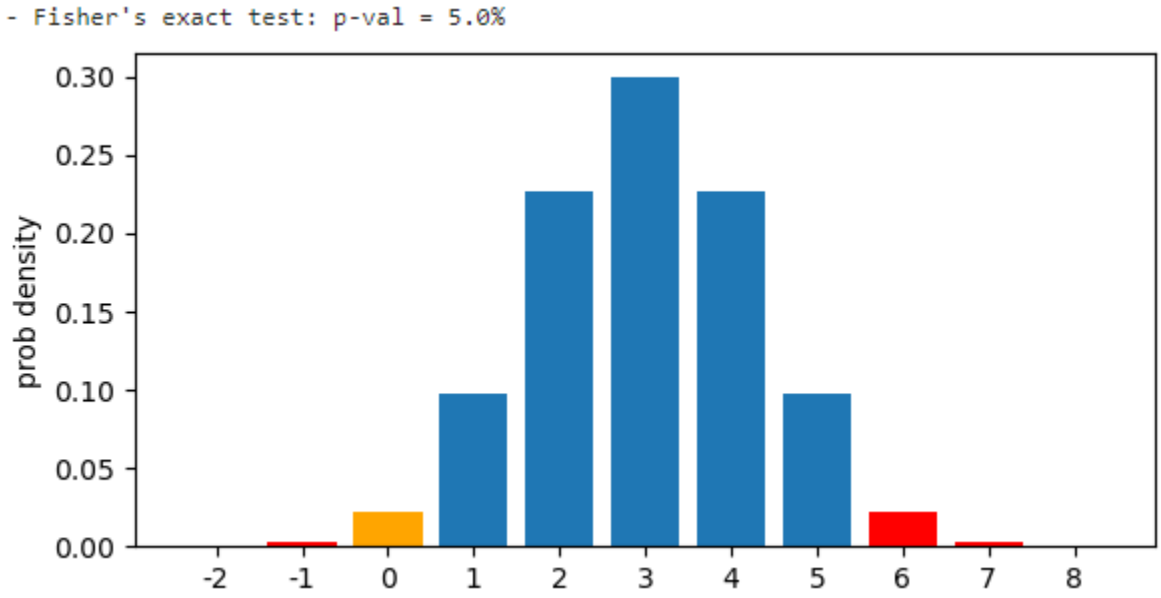

Fig.14 Represents the area under the two-tail test graph.

4.7 Continuous metrics and student t-test

The other parameter that is required in order to complete the test falls under the umbrella of continuous metrics which is given to us by the use of tasks the users were made to complete. Similar to the discrete metrics a unique question that is considered in this case is how long the user takes in order to complete all the assigned tasks.

Let X depict the users that utilized visualization model A and Y represent visualization model B. This can be observed in Fig.15. The two-dimensional data that is obtained will provide the statistical discrepancy when we use the student t-test.

```
- Observations:
  - Version A: = [200 150 250 350 150 150 350 250 150 250 150 150 200   0   0]
  - Version B: = [200 150 300 150 150 400 250 250 150 200 250 150 300 200 250]

- Distribution plot:
```

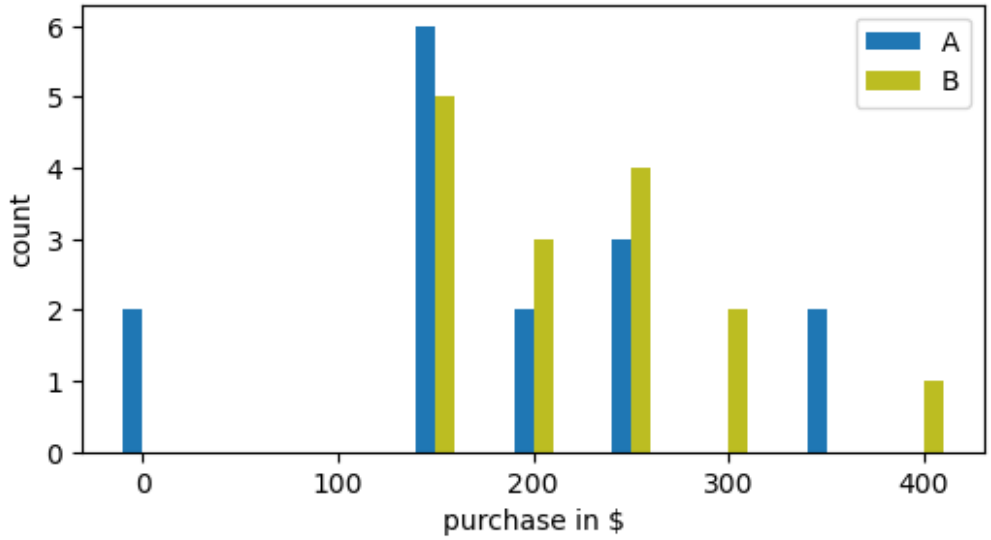

Fig.15 depicts the distribution plot between models A and B

Student's t-test [15] can then be applied under the following assumptions.

- The observations are normally distributed (or the sample size is large).
- The sampling distributions have "similar" variances σX ≈ σY.

Thus on assuming the above conditions, the student t-test relies on the t-value obtained from the formula in Fig.16.

$$T = \frac{\bar{X} - \bar{Y}}{S_P \cdot \sqrt{1/n_X + 1/n_Y}} \sim t_\nu \qquad \nu = n_X + n_Y - 2$$

Fig.16 Represents the student test formula

Here SP is the pooled standard deviation obtained from the sample variances SX and S Y, which are computed using the unbiased formula that applies Bessel's correction. Using these values the p-value can be determined.

Using the t-value obtained from the student t-test and the p-value obtained from the Fishers test it is possible to determine which hypothesis is correct with the use of the area under the graph in Fig.17.

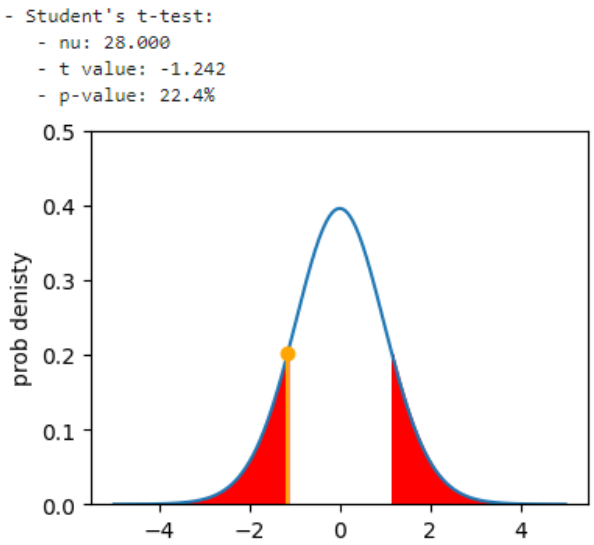

Fig.17 Represents the results obtained from the student t-test and the area under the graph.

4.8 Questionnaire

The questionnaire has five sections, consisting of subjective details and objective measures. The goal of this questionnaire is to provide an idea of the user standpoint of the project. The questions are meant to capture details about the user, model, and further changes to be made.

4.8.1 Pre-study questions

Before the testing process begins, it is paramount that the participants clear the basic criteria that are set beforehand. This process will act as a screening feature to weed out the participants that are not qualified to partake in the experiment.

- *Question 1*: Know the user (initial user information)
    - 1.1 Kindly provide a few details about yourself. What course are you Majoring in? And the year of study at Dalhousie.
    - 1.2 Describe your typical day at Dal as a _____. (with respect to your role)
    - 1.3 When you are on the computer, do you experience difficulty accessing information?

- *Question 2*: Gathering user behavior (user's needs with the application)
    - 2.1 What are the most important tasks you or other people need to perform in using the prototype?
    - 2.2a How would you describe your past experience with ensembles?
        - Very poor/Poor/Fair/Good/Excellent
    - 2.2b How would you describe your current experience with ensembles?
        - Very poor/Poor/Fair/Good/Excellent
    - 2.3 What devices do you typically use when coding?
    - 2.4 Do you or did you in the past use other websites and resources for the same purpose as the prototype? Yes/No

- 2.5 Is there anything you or your users often look for on the application that is missing or hard to find?
- 2.6 Is there any way application isn't supporting your needs currently?

4.8.2 Measure of Study Questions

In order to evaluate the tasks that are completed questions that answer the qualitative results of the user are required.

- *Question 3*: Quantitative results (time taken by users for the tasks)
  - 3.1 How long did the user take to complete task 1? (in mins)
  - 3.2 Identify all instances of mislabeling and disagreement? (Yes/No)

These questions help to determine the statistical disparity between the two different glyphs. Question 3.1 will serve as the core functionality of the continuous metrics that are detailed in section 4.7. And question 3.2 will serve as the core functionality of the discrete metrics that are detailed in section 4.6.

4.8.3 Post-Study Questions

These Questions serve the purpose of gathering user feedback and their impression of the website as a whole user experience.

- *Question 4*: Gathering opinion
  - 4.1 What do you see as the primary function of the application?
  - 4.2 What do you like about the current application?
  - 4.3 What don't you like about the current application?

- *Question 5*: Information provision
  - 5.1 How do you use the information on the application?
  - 5.2 Would you ever need to share these metrics with others?

- *Question 6*: Customization Questions (the ideal configuration that is requested by the majority of the users)
  - 6.1 Does the user prefer the polygonal glyph over the flower glyph?
  - 6.2 If the user prefers the polygonal glyph why?
  - 6.3 Any further changes made to the dashboard for better understanding?

4.8.4 Workload-based questions

A common universally used methodology to evaluate and understand the workload of the user is the NASA Task Load Index (TLX) method [16]. This assesses workload on a five-7-point scale. Increments of High, Medium, and low estimates for each point result in the 21 gradations on the scale. This can be seen in Fig.18.

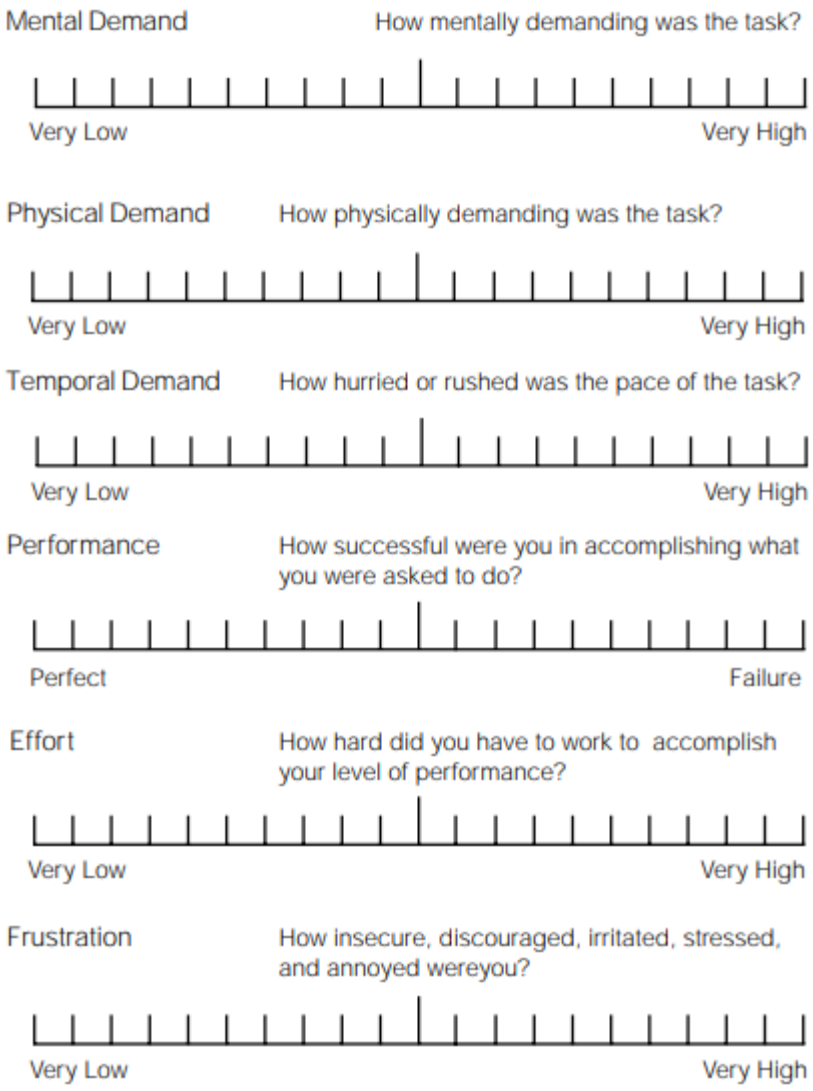

Fig. 18 Represents the NASA TLX form and the scale for evaluation

4.8.5 Final Remarks

In order to determine the validity of this questionnaire and to determine any changes to be made to this questionnaire, final remarks are taken into account.

- *Question 7*: Closing questions
    - 8.1 What haven't we asked you today that you think would be valuable for us?
    - 8.2 May I contact you if we have any other questions or for possible further research for this project?

Further changes could be performed after the review is complete.
Evaluation, future work, and conclusion will be completed after the user study is complete.

5.0 References

1. Luz, Christian F., et al. "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies." Clinical Microbiology and Infection 26.10 (2020). (pp 01-08)
2. Phadke, Madhura N., et al. "Exploring ensemble visualization." Visualization and Data Analysis 2012. Vol. 8294. SPIE, 2012. (pp 01-18)
3. De Winter, Joost CF. "Using the Student's t-test with extremely small sample sizes." Practical Assessment, Research, and Evaluation 18.1 (2013).
4. Kammer, Dietrich, et al. "Glyphboard: Visual exploration of high-dimensional data combining glyphs with dimensionality reduction." IEEE transactions on visualization and computer graphics 26.4 (2020). (pp 01-17)
5. Hilasaca, Gladys M., et al. "Overlap Removal of Dimensionality Reduction Scatterplot Layouts." arXiv preprint arXiv:1903.06262 (2019). (pp 02-17)
6. Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques." Artificial Intelligence Review 26.3 (2006). (pp 01-17)
7. Yang, Bo-Suk, Xiao Di, and Tian Han. "Random forests classifier for machine fault diagnosis." Journal of mechanical science and technology 22.9 (2008). (pp 01-21)
8. Mazumdar, Dipankar, Mário Popolin Neto, and Fernando V. Paulovich. "Random Forest Similarity Maps: A Scalable Visual Representation for Global and Local Interpretation." Electronics 10.22 (2021). (pp 06-24)
9. Flury, Bernhard, and Hans Riedwyl. "Graphical representation of multivariate data by means of asymmetrical faces." Journal of the American Statistical Association 76.376 (1981). (pp 93-121)
10. Bruckner, Stefan, and Meister Eduard Groller. Volumeshop: An interactive system for direct volume illustration. IEEE, 2005. (pp 671-678)
11. Amershi, Saleema, et al. "Guidelines for human-AI interaction." Proceedings of the 2019 chi conference on human factors in computing systems. 2019. (pp 01-14)
12. Cho, Hyun-Chul, and Shuzo Abe. "Is two-tailed testing for directional research hypotheses tests legitimate?." Journal of Business Research 66.9 (2013). (pp 01-22)
13. Connelly, Lynne M. "Fisher's exact test." MedSurg Nursing 25.1 (2016). (pp 01-04)
14. Hoonakker, Peter, et al. "Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX)." IIE transactions on healthcare systems engineering 1.2 (2011). (pp 05-13)
15. Amershi, Saleema, et al. "Guidelines for human-AI interaction." Proceedings of the 2019 chi conference on human factors in computing systems. 2019. (pp 01-14)
16. Rechkemmer, Amy, and Ming Yin. "When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models." CHI Conference on Human Factors in Computing Systems. 2022. (pp 01-16)