RCL Mastery of Tabular Data:
Outperforming Conventional Models in Stability and Accuracy

Eslam Ahmed Abdelrahman*

ABSTRACT

Abstract— In collaboration with Lumina AI, this research discusses the effectiveness of Random Contrast Learning (Lumina RCL) on a variety of tabular datasets in comparison to more traditional methods of machine learning. It describes how well Lumina RCL handles small imbalanced datasets, resists overfitting, and generalizes without need for balancing data, feature selection, or feature scaling.

I. INTRODUCTION

Despite recent rapid advancements in the world of machine learning, it remains challenging to find models that can learn effectively from limited and complex data. In general, traditional machine learning and deep learning models, although powerful, often struggle with small or imbalanced datasets. Without vast amounts of data or computing power, resulting models are often inefficient or even inapplicable for many real-world problems.

Random Contrast Learning (Lumina RCL) is a groundbreaking algorithm created to succeed in places where traditional models fail. Lumina RCL uses randomness as a filter to identify precisely which variables constitute patterns of interest. This technique, modeled after human consciousness, allows Lumina RCL to discern the most salient features of an entire dataset and to detect patterns even when datasets are small. Regardless of the size of datasets, Lumina RCL avoids becoming lost or misled by irrelevant details, never losing the forest for the trees. This represents a significant departure from traditional models that analyze data meticulously, piece by piece. Moreover, having organized the dataset, Lumina RCL retains the data with near perfect recall. In other words, Lumina RCL both learns and memorizes. Lumina RCL processes data with both efficiency and profundity, needing less data, requiring less compute, and without any sacrifice of accuracy.

Previously, two research papers explored the use of Lumina RCL in different applications of image and text classification. “Redefining Words: The Power of RCL in Text Classification” emphasized the tool's ability to manage duplicate data efficiently without loss of insight. “Binary Image Classification Comparison Using Neural Networks and Random Contrast Learning” showcased Lumina RCL’s exceptional performance in medical image classification, where it displayed remarkable efficiency and accuracy even with minimal data. This paper continues to elaborate on just the beginnings of Lumina RCL’s capabilities. This paper covers the performance of Lumina RCL vs. traditional models on tabular data, specifically considering performance with unbalanced data, resistance to overfitting, and potential effects of dataset size and feature selection or scaling processes.

* Author Eslam is a Machine Learning Engineer, an AI and Machine Learning Instructor at more than one institution, e.g. Udacity, and an ML author at Educative. He holds two master’s degrees in AI and ML from the University of Ottawa and IoT from Benha University. He is certified as a machine learning engineer and data scientist from Microsoft and IBM. Email: eslamahme@gmail.com

II. DESCRIPTION OF DATASETS

The datasets chosen for this research are public datasets used for benchmarking and evaluating machine learning models and represent a diverse set of challenges, especially in relation to small datasets across a wide array of applications. These datasets, both balanced and imbalanced, pertain to medical diagnostics, industrial categorization, and agricultural classification, call for binary classification and multi-class classification, and exhibit differences in complexity and forms of data distribution.

1. **Pima Indians Diabetes Database** (768 samples, 8 attributes, 2-class, Unbalanced)
2. **Iris Flower Dataset** (150 samples, 4 attributes, 3-class, Balanced)
3. **Wheat-Seeds Dataset** (210 samples, 7 attributes, 3-class, Balanced)
4. **Glass Identification Dataset** (214 samples, 9 attributes, 6-class, Unbalanced)
5. **Haberman's Survival Dataset** (306 samples, 3 attributes, 2-class, Unbalanced)

2.1 Pima Indians Diabetes:

This data collection includes measurements used to predict the likelihood of diabetes, in patients of Pima Indian descent aged 21 and older. It covers factors like pregnancies, glucose concentration, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age.

2.2 Iris Flower:

This classic dataset in machine learning and statistics, the Iris Flower Dataset, is commonly used for classifying three species of iris flowers (Setosa, Versicolor, and Virginica) based on attributes like sepal length, sepal width, petal length, and petal width.

2.3 Wheat-Seeds:

This dataset includes characteristics of kernels from three distinct types of wheat. The features measured include area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove.

2.4 Glass Identification:

This dataset is designed for categorizing glass types based on their chemical composition. It is often utilized in forensic science investigations to aid in analyzing glass samples found at crime scenes. The attributes considered are the index and the proportions of the eight elements found in the glass (Na, Mg, Al, Si, K, Ca, Ba, and Fe).

2.5 Haberman's Survival:

This dataset contains information on survival outcomes for patients who underwent breast cancer surgery. The factors taken into consideration are the age of the patient during the surgery, the year when the surgery took place, and the count of nodes found, in addition to whether the patient survived for five years after the operation.

III. METHODOLOGY

We conducted 423 experiments to compare Random Contrast Learning (Lumina RCL) with traditional machine learning (ML) and deep learning (DL) models. These include training on the original datasets, analyses of overfitting, and variations with undersampling, oversampling, feature selection and feature scaling. We applied Lumina RCL and non-RCL experiments to each dataset several times and will only present in the next section the best sets of results achieved. One hundred seventy of the experiments included Lumina RCL. 232 experiments were conducted to assess eight different ML models. Twenty-one experiments were exclusively designed for DL models. Before considering the experiments themselves, this section begins by reviewing the specifications for the use of Lumina RCL, the ML models, and the DL models, respectively.

3.1 Lumina RCL:

Lumina RCL produces models that achieve better inference and exhibit compact size. As mentioned above, this is made possible by Lumina RCL's use of randomness as a filter to identify patterns in tabular data. Consequently, there is less need for very large datasets. Lumina RCL automatically fine-tunes its models to the specific characteristics of the data to achieve high performance with far fewer computational resources. To do this, Lumina RCL uses an auto-optimize parameter in training – one of its most significant features. This enables Lumina RCL to learn accurately from tabular data.

In our tabular dataset experiments, we used the Lumina RCL Classifier algorithm (RCLC) via the PrismRCL application. We ran the application on Windows machines powered by Intel i7 processors with 128 GB of RAM and DDR4. The model setup is straightforward, allowing optimization through a single command line. Lumina RCL is designed to work with full CPU power in a multi-threaded application environment.

3.2 Machine Learning Models:

For our tabular dataset experiments, we carefully selected a set of machine learning models, considering their unique advantages and diverse architectures. These ML models show how traditional models perform on tabular data and therefore allow us to showcase the innovative capabilities of Lumina RCL. These experiments have been performed on the same hardware as the Lumina RCL experiments: Intel i7 Processor with 128 GB DDR4 RAM, and CPU processing without GPU.

- **Logistic Regression** has been included as a test against all other models due to its simple form and applicability to most binary classification tasks.
- **Decision Tree and Random Forest** were selected for their simplicity and enhanced accuracy, respectively, effectively addressing diverse dataset challenges.
- **Gradient Boosting and XGBoost** were selected for their high performance with sequential error correction and efficient handling of all types of data.
- **SVM (Support Vector Machine)** was selected for its ability to establish complex class boundaries.
- **KNN (K-Nearest Neighbors)** was selected for its ability to classify based on similarity.

- **Naive Bayes** has been included for its performance with datasets exhibiting feature independence.

3.3 Deep Learning Models:

For our tabular dataset experiments, we developed a unique neural network on Google CoLab for each dataset based on its features and complexities. For each dataset, we experimented in TensorFlow with activation functions like sigmoid, SoftMax, and ReLU, with an Adam optimizer, to get the best fit for the characteristics of the dataset. We performed the tests and fine-tuned our models to identify the most suitable configuration in terms of performance to be used with each neural network. These experiments have been performed on hardware with the following specifications:

- **GPU:** 1xTesla T4, with 2560 CUDA cores, compute 3.7, and 15GB (15.079GB Usable) GDDR6 VRAM.
- **CPU:** 1xsingle-core hyper-threaded Xeon Processor @ 2.3GHz (No Turbo Boost) with 45MB Cache.
- **RAM:** Approximately 12.7GB available.
- **Disk:** Approximately 78GB available.

3.4 Comparative Experiments

3.4.1 Original Dataset Experiment

This initial experiment used all five of the original datasets without any modifications. The objective was to assess the ability of Lumina RCL, the selected ML models, and the DL models to handle data in its unaltered form. By avoiding data shaping techniques like balancing or feature selection, we aimed to measure the capacity of each model to understand and classify data as-is. This approach allowed us to establish a baseline performance metric for all models, and to highlight their innate strengths and limitations in processing raw data.

3.4.2 Overfitting Analysis

This experiment was designed to evaluate how well Lumina RCL, the selected ML models, and the deep learning models to avoid overfitting with small or unbalanced datasets across five data scenarios. This issue arises when a model learns too much from its training data and is measured by comparing how it performs on new, unseen data. This experiment determines whether each model merely memorizes patterns or achieves an understanding of them. The latter, which we may also understand as resistance to overfitting, is demonstrated by a small difference in accuracy between training and testing.

3.4.3 Balanced vs. Unbalanced Data Handling (Undersampling and Oversampling):

Because the **Iris wheat** and **Wheat-Seed** datasets are already balanced, we next focused on the unbalanced **Pima Indians Diabetes, Glass Identification, and Haberman's Survival** datasets**.** These are small datasets, so even a little imbalance could potentially sway our results. After balancing these datasets through undersampling or oversampling we tested the accuracy of Lumina RCL, the selected ML models, and the deep learning models, and compared these accuracies with our initial experiment.

- **Undersampling Experiment**: Undersampling reduces the size of larger classes to match the smaller classes. This technique is rarely applied to small datasets. A significant loss of model accuracy indicates the loss of important information, i.e. the model no longer fully understands the data, which is also known as underfitting.
- **Oversampling Experiment**: Oversampling increases the size of smaller classes to match the larger classes. Improvement in accuracy after oversampling indicates that increasing the minority class size helped the model learn better about underrepresented classes, but oversampling can lead to overfitting, i.e. the model memorizes the oversampled data instead of learning from it.

3.4.4 Feature Selection

Traditionally, feature selection is a critical step in machine learning. We sought to determine whether feature selection improves the performance of Lumina RCL, in contrast with traditional models. For this experiment we used **Wheat-Seeds**, **Glass Identification, and Haberman's Survival** datasets to test across balanced and unbalanced, binary, and multiclass datasets. Because these datasets have numerical input data and a categorical target variable, we used for feature selection ANOVA (Analysis of Variance) and wrapper methods. We compared the rate of accuracy with and without the use of feature selection.

3.4.5 Feature Scaling

Especially for logistic regression, other non-tree-based algorithms, and some deep learning models, feature scaling is an important preprocessing step that standardizes the variable range to ensure that all features contribute equally to model training. We sought to determine whether the accuracy and performance of Lumina RCL remained consistent after feature scaling, in contrast with other models that are usually impacted by this preprocessing step.

IV. RESULTS

This section presents the comprehensive results of our experiments comparing the performance of Lumina RCL against traditional machine learning and deep learning models across the five datasets. The results demonstrate Lumina RCL's stability and accuracy with imbalanced data and its resistance to overfitting. They also showcase Lumina RCL's effectiveness with small datasets and its independence from feature selection and feature scaling processes. Lumina RCL provides metrics for recall, f1-score, accuracy, precision for each class, and overall accuracy for all classes. Here we will concern ourselves with accuracy.

4.1 Original Dataset Experiment Results:

We began by testing the accuracy of Lumina RCL, the selected ML models, and the deep learning models on all five datasets without any modifications, as displayed in Table 4.1 and Figure 4.1.

Dataset	RCL Model	DL model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XG Boost	SVM	KNN	Naïve Bayes
Pima Indians	74.10	72.73	77.59	71.09	75.77	76.42	74.47	75.77	69.92	75.12
Iris	95.70	96.67	97.33	94.00	96.00	95.33	94.0	96.6	97.33	96.00
Wheat-seeds	93.80	87.14	92.38	87.61	92.38	92.85	93.33	90.95	89.04	90.95
Glass	70.00	58.14	62.79	66.69	77.03	74.70	77.01	33.27	66.17	55.28
Haberman	72.70	72.59	72.84	65.99	67.67	70.78	64.20	71.45	72.82	74.24

Table 4.1. Accuracy values for Lumina RCL, deep learning, and machine learning models on all datasets.

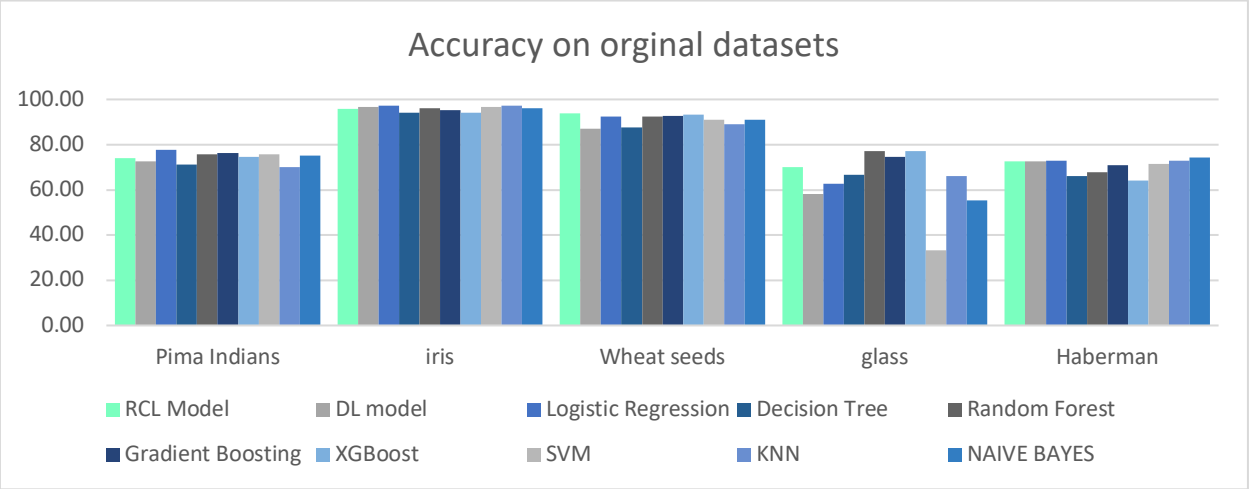

Figure 4.1. Accuracy values for Lumina RCL, deep learning, and machine learning models on all datasets.

As expected, the range of models tested yielded a range of results. Noteworthy are the poor performances of the DL model (58.14%), SVM (33.27%), and Naïve Bayes (55.28%) on the **Glass Identification** dataset. These results provided a baseline for further evaluation. These results provided a baseline for further experimentation under various data transformations.

4.2 Overfitting Analysis Results:

Next, we tested how well Lumina RCL, the selected ML models, and the DL models learned rather than memorized patterns in the datasets. For some of the models, we found significant changes between training and testing accuracies. This indicates a widespread problem of overfitting. In other words, these models memorized instead of learned; they failed to understand the data. Table 4.2 shows the delta values between train set accuracies and test set accuracies for each of the models and each of the five datasets.

$$delta\ value = \ train\ set\ accuracy\ value - test\ set\ accuracy\ value$$

Dataset	RCL Model	DL model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XG Boost	SVM	KNN	Naïve Bayes
Pima Indians	2.4	1.7	1.65	29.31	25	10.78	22.41	4.18	13.84	1.23
Iris	0.6	0.33	1.99	8.7	8.7	4.35	8.7	6.33	3.94	5.55
Wheat-seeds	0.9	2.03	0.95	3.13	6.25	3.13	3.13	-3.3	4.07	0.39
Glass	1.4	1.99	3.8	31.25	21.88	21.88	28.13	-2.14	13.19	9.48
Haberman's survival	0.3	3.89	1.64	41.13	32.04	18.48	31.22	-0.19	11.33	1.64
Average difference	**1.12**	**1.988**	**2.006**	**22.704**	**18.774**	**11.724**	**18.718**	**3.228**	**9.274**	**3.658**

Table 4.2. Delta values between the training accuracies and test accuracies.

As shown in Table 4.2, Lumina RCL exhibited the lowest delta values for the **Pima Indians, Glass Identification, and Haberman's Survival** datasets - 2.4, 1.4, 0.3, respectively. This clearly contrasts with double-digit delta values, such as of the Decision Tree and Random Forest models, which indicate significant overfitting. Lumina RCL stands out when we compare the averages delta values across all five datasets for each of the tested models. For a better illustration, see the average delta value across all datasets for each model as shown in Figure 4.2.

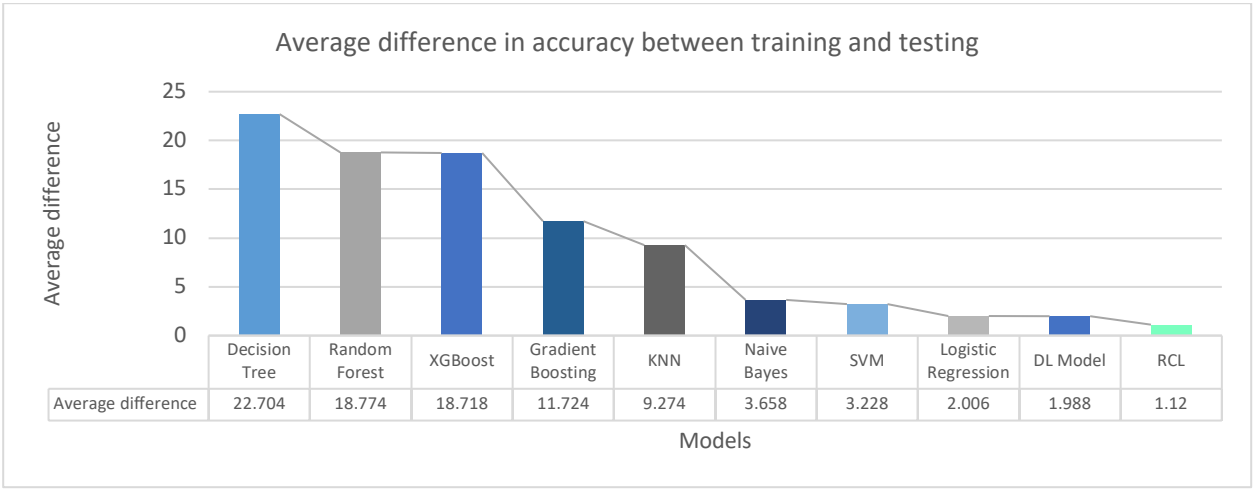

Figure 4.2. Average delta values between training accuracies and test accuracies from greatest to least.

The results presented in the table and the figure above show the stability of Lumina's Random Contrast Learning (Lumina RCL) against overfitting. Lumina RCL has an average difference between training and testing accuracies of 1.12 - significantly less than most traditional models.

4.3 Balanced vs. Unbalanced Data Handling Results:

Real-world data is often uneven, which can cause bias toward prevalent classes. The next set of experiments measured how balancing datasets through undersampling and oversampling affected model accuracies.

4.3.1 Undersampling Experiment Results:

As mentioned before, undersampling reduces the size of larger classes to match smaller ones. This method risks losing information in the dataset and thereby decreasing model performance. Undersampling is typically not applied to small datasets, but we tested this anyway to determine its effect on Lumina RCL. Table 4.3.1 shows the delta values for each model:

$$delta\ value = |original\ data\ accuracy\ value - undersampled\ data\ accuracy\ value|$$

Dataset	RCL Model	DL model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XG Boost	SVM	KNN	Naïve Bayes
Pima Indians	0.2	9.74	2.86	4.94	3.77	1.94	1.69	3.24	1.95	1.42
Glass	0.1	13.95	15.19	19.33	10.43	25.46	28.22	0.41	19.65	14.59
Haberman's survival	2.3	11.42	4.5	10.78	5.97	11.41	5.51	0.69	11.75	4.89
Average difference	0.87	11.70	7.517	11.68	6.72	12.94	11.81	1.45	11.12	6.97

Table 4.3.1 Delta values between the original dataset accuracies and accuracies after undersampling.

Among all models, Lumina RCL had the lowest average delta value at just 0.87. This suggests that Lumina RCL does not require data balancing and readily recognizes minority classes in naturally unbalanced datasets. As expected, when applying undersampling to small datasets, many models saw significant drops in accuracy. (Note: Table 4.3.1 reflects only absolute values.) Especially with the **Glass Identification** dataset, most competitors suffered even double-digit losses in accuracy. While SVM showed a very small delta value, SVM did not exhibit an impressive original dataset accuracy to begin with, as shown in both Table 4.1 and Figure 4.1. These results illustrate RCL's remarkable ability, compared with other models applied to small datasets, to maintain performance even when the data is undersampled.

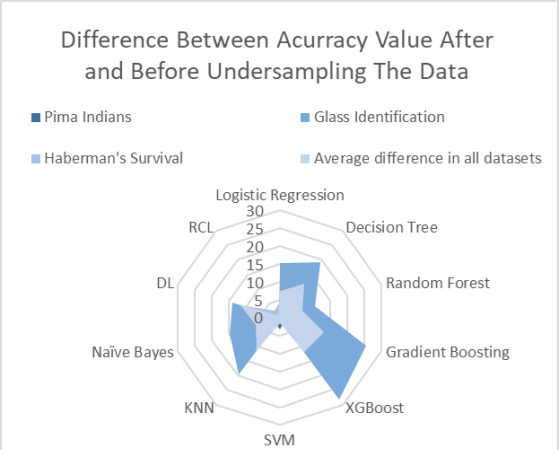

Figure 4.3.1 Differences in accuracy values before and after undersampling.

4.3.2 Oversampling Experiment Results:

Opposite of undersampling, oversampling resolves imbalances in datasets by increasing the size of the smaller classes through repetition or interpolation to match the larger classes. This technique may improve model performance, but it risks overfitting. To mirror our undersampling experiment, we applied oversampling to the whole datasets to measure its effect on the accuracy of Lumina RCL, the selected ML models, and the DL models. Table 4.3.2 shows the delta value for each model:

$$delta\ value = |original\ data\ accuracy\ value - oversampled\ data\ accuracy\ value|$$

Dataset	RCL Model	DL model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XG Boost	SVM	KNN	Naïve Bayes
Pima Indians	0.60	2.60	3.89	1.56	4.56	0.52	0.90	3.64	4.55	0.38
Glass	0.20	6.98	5.65	1.95	14.05	2.27	1.34	1.06	1.43	19.15
Haberman's survival	0.30	2.79	2.42	1.78	0.76	4.19	1.38	2.39	14.16	0.03
Average difference	**0.37**	**4.12**	**3.99**	**1.76**	**6.46**	**2.33**	**1.21**	**2.36**	**6.71**	**6.52**

Table 4.3.2 Delta values between original accuracies and accuracies after oversampling.

The results were varied. In most cases, model performance was negatively impacted by oversampling, suggesting that the technique does not work well with small datasets. (Note: Table 4.3.1 reflects only absolute values.) However, the delta value for Random Forest on the Glass Identification dataset was an improvement. Just as in the undersampling experiment, Lumina RCL showed the lowest average delta value at 0.37. Again, this suggests that Lumina RCL readily recognizes classes in naturally unbalanced datasets and even showcases its robustness in the face of manipulated datasets. The stability of Lumina RCL, regardless of the dataset type, demonstrates that its ability to generalize from data neither needs nor suffers due to oversampling. This feature is very useful in real-life scenarios where data imbalance is a common issue. For a better illustration, see the average delta value for each model in all datasets in Figure 4.3.2.

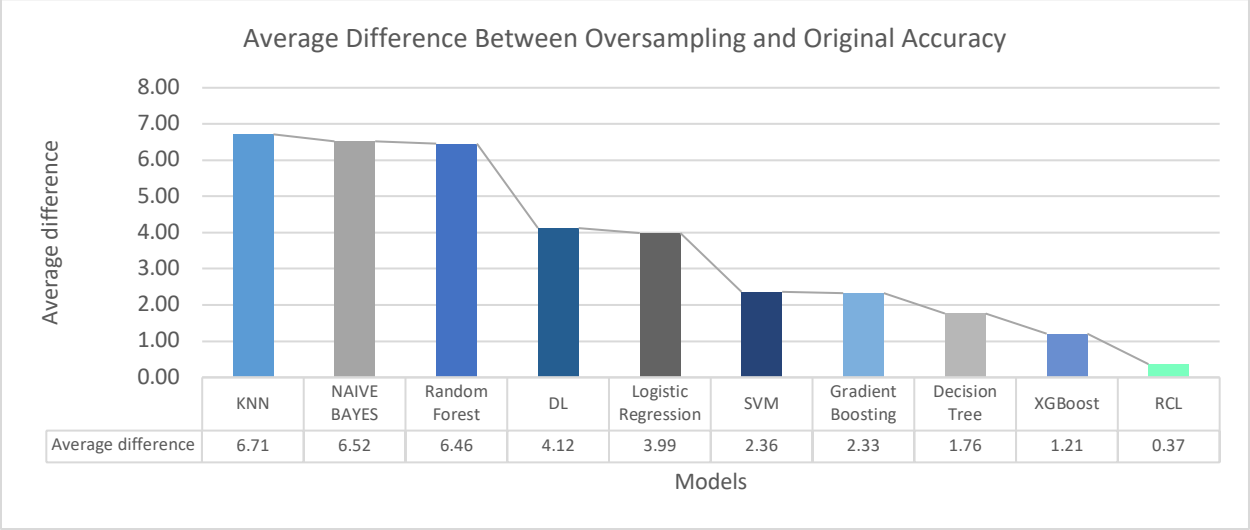

Figure 4.3.2 Average delta values after applying Oversampling on all models from greatest to least.

Because the effect on the accuracies of the models by oversampling were varied, we decided not to proceed with testing consequent overfitting.

4.4 Feature Selection Experiment Results:

Next, we tested the effects of feature selection on Lumina RCL, the selected ML models, and DL models. Most models show modest gains in accuracy. ~~Feature Selection Experiment Results:~~In Table 4.4, we determined the delta value for each model:

$$delta\ value = |\ original\ data\ accuracy\ value\ -\ dataset\ accuracy\ value\ after\ feature\ selection|$$

Dataset	RCL Model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XGBoost	SVM	KNN	Naïve Bayes
Wheat-seeds	0.81	1.91	3.34	3.33	2.37	5.23	1.9	2.39	1.91
Glass	0.62	0.93	0.91	1.84	3.22	0.56	0	2.32	15.02
Haberman's survival	0.75	0.35	6.85	7.6	0	9.69	1.41	0.31	2.77
Average difference	**0.74**	**1.063**	**3.7**	**4.26**	**1.86**	**5.16**	**1.10**	**1.67**	**6.567**

Table 4.4 Delta values between original accuracies and accuracies after feature selection.

Lumina RCL shows the smallest average delta value among all models tested. This suggests that feature selection provides no benefit beyond the native feature-selecting capacity of Lumina RCL. This is displayed in Figure 4.4.

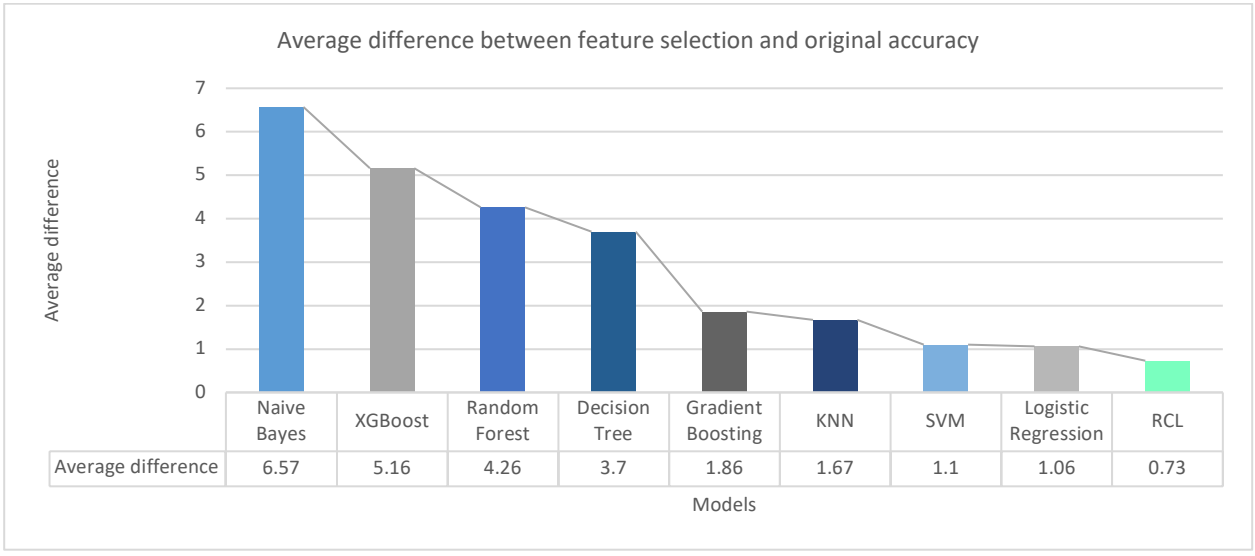

Figure 4.4 Average delta values after applying Feature Selection on all models from greatest to least.

~~4.4~~4.5 Feature Scaling Experiment Results:

Feature scaling is a technique that standardizes the range of features in a dataset. Table 4.5 shows how each model responds to this technique:

$$delta\ value = |original\ data\ accuracy\ value - dataset\ accuracy\ value\ after\ feature\ scaling|$$

Dataset	RCL Model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XG Boost	SVM	KNN	Naïve Bayes
Pima Indians	0.3	0.51	0.01	0.13	0	0.01	0.4	2.99	0
Iris	0.28	2	0	0	0.66	0	0.6	2	0
Wheat-seeds	0.2	1.43	0.01	0.48	0.01	0	2.38	4.29	0
Glass	0.47	1.43	0.46	0.94	0.46	0	37.6	0.5	12.12
Haberman's survival	0.03	0.34	0.02	1.04	0	0	1.05	2.07	0.35
Average difference	**0.256**	**1.142**	**0.1**	**0.518**	**0.226**	**0.002**	**8.406**	**2.37**	**2.494**

Table 4.5 Delta values between original accuracies and accuracies after feature scaling.

As shown in Table 4.5 Lumina RCL exhibited little change in accuracy due to feature scaling. Other tree-based models – Decision Tree, Random Forest, Gradient Boosting, and XGBoost – behaved similarly. Consequently, insensitivity to feature scaling processes is a notable characteristic of tree-based models generally rather than Lumina RCL specifically. See their positions in Figure 4.5.

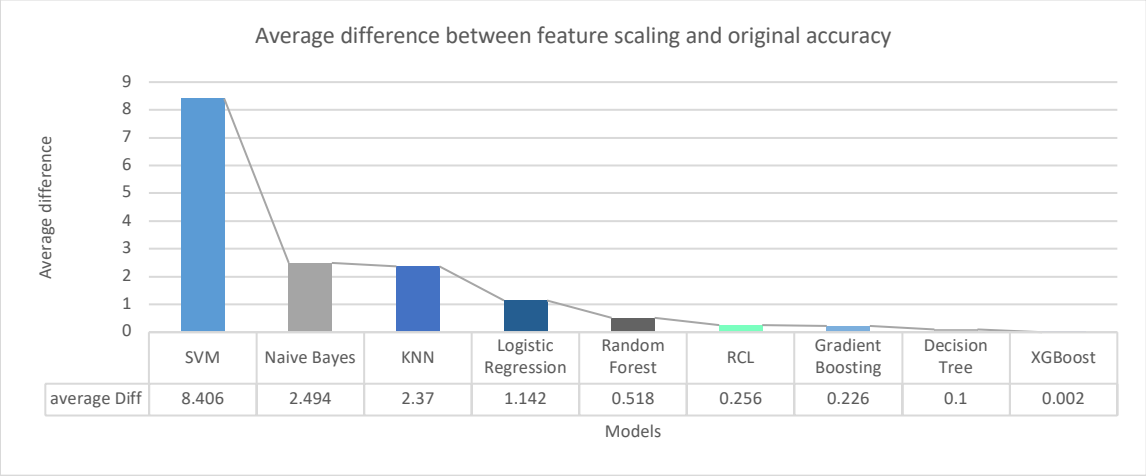

Figure 4.5 Average delta values after applying feature scaling on all models from greatest to least.

Overall Comparative Analysis:

Machine learning models respond to different data transformations in different ways. The average delta values across the relevant datasets before and after the applications of data transformation are summarized in Table 5.1.

Average difference	RCL	DL models	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XG Boost	SVM	KNN	Naïve Bayes
overfitting	1.12	1.988	2.006	22.704	18.774	11.724	18.72	3.23	9.274	3.658
undersampling	0.58	7.98	4.832	7.392	5.444	8.088	7.084	1.09	6.766	4.18
oversampling	0.28	3.33	2.62	1.16	5.32	1.81	0.72	1.43	4.16	3.91
feature selection	0.73	-	1.06	3.7	4.26	1.86	5.16	1.1	1.67	6.57
feature scaling	0.26	-	1.142	0.1	0.518	0.226	0.002	8.41	2.37	2.494

Table 5.1 The average delta values of dataset accuracies for each experiment before and after data transformation.

In four of five experiments, Lumina RCL maintains the lowest average delta value. In the feature scaling experiment, RCL exhibited behavior similar to other tree-based models. This stability suggests a strong capacity for generalization, i.e. an understanding of the underlying patterns in data, despite difficulties with small or imbalanced datasets. In contrast, the performance of other models is dependent on the selection of data transformation techniques. Based on these results, we believe that Lumina RCL fills a gap in the world of machine learning. It learns effectively from limited and complex data. It resists overfitting. It readily recognizes underrepresented classes. It is insensitive to oversampling. And it requires neither feature selection nor feature scaling but handles data well in its natural form. These are precisely the characteristics of human consciousness that Lumina RCL was designed to mimic.

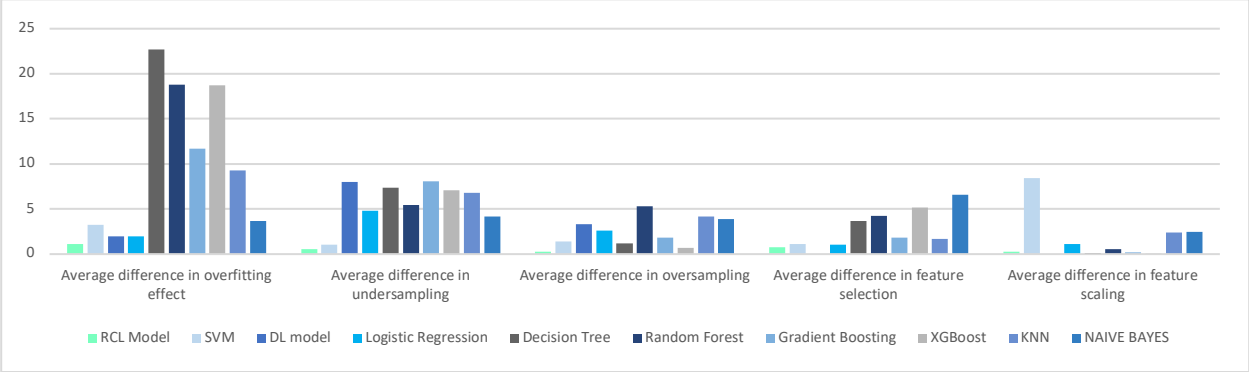

Figure 5.1 Average delta values of dataset accuracies for each experiment before and after data transformation.

V. CONCLUSION

We conducted 521 experiments on five public datasets widely used for benchmarking, simulating a diverse set of challenges with small datasets. We tested Lumina RCL, eight different machine learning models, and five deep learning neural networks – one for each of the datasets. We then tested all of these with five different data transformations. We found that Lumina RCL, in comparison with all other models, exhibits remarkable stability and accuracy with datasets that are small and naturally imbalanced. Lumina RCL was designed to mimic human consciousness and for this reason excels with limited and complex data typically found in real-world situations. We have demonstrated that Lumina RCL neither needs nor suffers from preprocessing steps of data transformation or manipulation to maintain its accuracy. Moreover, in addition to these characteristics, the cost-savings of Lumina RCL, which runs on CPU, holds immense potential for the entire field of machine learning.