Simple Models, Solid Results: Linear/Logistic Regression with Mini-Batch SGD

COMP551 A1 Report

Nicholas Milin (261 106 314) – Diana Covaci (261 086 280) – Carl-Elliott Bilodeau-Savaria (261 155 640)

1 Abstract

This project investigates the performance of linear regression and logistic regression, two fundamental and widely used machine learning algorithms, on two biomedical benchmark datasets: the *Parkinson's Telemonitoring* dataset for regression of motor scores, and the *Wisconsin Breast Cancer Diagnostic* dataset for binary tumor classification. We compare standard analytical and full-batch gradient descent solutions with a mini-batch stochastic gradient descent (SGD) implementation to assess its potential for scalability and improved generalizations. A key finding is that a 70/30 train/test split consistently yields the highest performance, underscoring the importance of sufficient training data. Notably, on the Parkinson's dataset, mini-batch stochastic gradient descent outperformed analytical linear regression for predicting the Total UPDRS score on the test set, highlighting its practical advantage in reducing overfitting. Overall, our experiments demonstrate the effectiveness and applicability of these models to biomedical prediction tasks and highlight how factors like training set size, batch size, and optimization technique significantly influence predictive outcomes.

2 Introduction

This project evaluates the efficiency of simple machine-learning models on two real-life datasets. We study 1. the *Parkinson's Telemonitoring* dataset, which consists of 5,875 at-home voice recordings from 42 patients and provides two clinician-rated targets (Motor and Total UPDRS) to be predicted from 19 features (16 of which are voice measures, with 3 being non-voice features) and 2. the *Wisconsin Breast Cancer Diagnostic* dataset, containing 569 biopsies, each described by 30 features (computed from digitized FNA images) and a binary target (M for a malignant tumor and B for a benign one).

Methodology. Our approach is to use linear regression (analytical form and mini-batch gradient descent) on the Parkinson's dataset since the UPDRS targets are continous numbers ranging from 0 to 108 and 0 to 176. As for the Breast Cancer dataset, we use logistic regression (with mini-batch SGD) since it's more apt for binary targets.

Key Findings. Our linear regression model on the Parkinson's dataset achieved test $MSE \approx 28$ for motor_UPDRS and ≈ 46.7 for total_UPDRS, with test $R^2 \approx 0.12 - 0.18$. We also found that mini-batch gradient descent outperformed the closed-form solution, which emphasizes its usefulness in preventing overfitting. On the Breast Cancer dataset, our logistic regression model attained high test accuracy, which is described later in this paper.

Background and related work. Both of these datasets have been extensively studied in machine learning literature. For the Parkinson's dataset, linear regression has already been tried with promising results, nearing clinicians' observations within 6 points for Motor UPDRS (out of 108) and within 7.5 points for Total UPDRS (out of 176) [7]. More advanced machine-learning models have also been tried (such as SVM, MLPNN, and GRNN) for this dataset, resulting in even more impressive results [3]. For the Breast Cancer dataset, logistic regression has also wielded impressive results, with a success rate as high as 96.51% [5]. Other models have also been tried such as MLP, Nearest Neighbor, Softmax Regression, and SVM, all resulting in a success rate over 90% [1].

3 Datasets

3.1 Parkinson Telemonitoring Dataset

The Parkinson's Telemonitoring dataset (UCI ID 189) contains 5,875 voice recordings from 42 patients with early-stage Parkinson's disease, featuring 19 biomedical measures and two target variables: motor_UPDRS and

total_UPDRS [7]. These recordings capture voice characteristics relevant to motor symptoms and were collected remotely in a telemonitoring study. The dataset contained no missing values or duplicate entries. Data preprocessing involved dropping the non-predictive subject ID column and scaling all remaining features using Scikit-learn's StandardScaler. Exploratory analysis of the dataset revealed correlations, particularly the five-jitter and six-shimmer measures, which quantify distinct aspects of vocal instability and amplitude variation. We retained all features as each captures unique, complementary information predictive of disease severity; removing any could result in a loss of valuable information and reduce model performance. Ethical considerations include the relatively small sample size and the male-biased cohort (28 males vs. 14 females), potentially limiting the generalizability of findings. Additionally, the remote collection method raises inherent concerns regarding patient privacy and data security [2].

3.2 Breast Cancer Diagnosis Dataset

The Breast Cancer Diagnostic dataset (UCI ID 17) comprises 569 instances, each representing a fine needle aspirate (FNA) of a breast mass, with 30 real-valued features derived from digitized images of cell nuclei. These features include mean, standard error, and "worst" (largest mean) values of various cell characteristics, such as radius, texture, perimeter, smoothness, symmetry, and fractal dimensions [6]. The dataset contained no missing values or duplicate entries. Preprocessing involved dropping the non-predictive sample ID column, scaling all features using Scikit-learn's StandardScaler, and encoding the target variable 'diagnosis' as a binary target (0 = benign, 1 = malignant). Exploratory analysis revealed a relatively balanced class distribution (357 benign, 212 malignant). Ethical considerations include the use of sensitive medical imaging data, which requires adherence to privacy regulations and robust data security measures to maintain patient confidentiality [4].

4 Results

4.1 Performance of Linear Regression and Logistic Regression

Table 1 presents the results of the closed-form linear regression with bias on an 80/20 train/test split of the Parkinson's Telemonitoring Dataset. The high mean squared errors (MSE) for motor_UPDRS and total_UPDRS, coupled with low R^2 scores (0.18), suggest that the underlying data may not follow a strong linear trend.

In contrast, Table 2 shows the performance of logistic regression on the Breast Cancer Diagnostic dataset using an 80/20 train/test split. The model achieved low cost values with minimal gap between training and test performance, indicating a good fit without significant overfitting.

Target Variable	Statistic	Train	Test
motor_UPDRS	**MSE**	27.866	28.007
	R^2	0.164	0.122
total_UPDRS	**MSE**	47.311	46.653
	R^2	0.180	0.158

Table 1: Performance of Linear Regression

Diagnostic Prediction	Cost
Training Set	0.043
Test Set	0.062

Table 2: Performance of Logistic Regression

4.2 Weights of Each Feature

For the analytical linear regression model, a majority of fitted coefficients are small in magnitude (approximately in the range -3 to 3), indicating a low linear impact. However, there are prominent outliers; Table 3 lists the largest coefficients for each target. Due to their large, oppositely signed weights, it is likely that Jitter:RAP and Jitter:DDP are collinear as well as Shimmer:APQ3 and Shimmer:DDA, a strong indicator of potential overfitting. These features have the most significant impact on the prediction of the target.

Target	Jitter:RAP	Jitter:DDP	Shimmer:APQ3	Shimmer:DDA
motor_UPDRS	-70.99	72.91	-197.63	195.53
total_UPDRS	-64.46	68.08	-129.82	126.87

Table 3: Weights of Features from Analytical Linear Regression

In contrast, the logistic regression model applied to the Breast Cancer Diagnostic dataset exhibited all relatively small feature weights (ranging from -1.26 to 2.30) with no outliers. The five heaviest positive weights are shown in Table 4. This balanced weight distribution indicates that no single feature dominates the prediction, contributing to the model's stable performance and reduced over-reliance on any single feature.

Feature	**radius2**	**texture3**	**symmetry3**	**concave_points1**	**area2**
Weight	2.302	2.250	2.103	1.847	1.659

Table 4: Top 5 Heaviest Weights from Logistic Regression

4.3 Effect of Increasing Subsets of Training Data on Performance

For both datasets, model performance peaked with a 70/30 train/test split when training each model on increasing subsets of data (20% to 80% in increments of 10%). For the linear regression model trained on the Parkinson dataset (Figure 1), the test MSE for both UPDRS targets was minimized at this ratio. Similarly, for the logistic regression model trained on the Breast Cancer dataset (Figure 2), test cost reached its minimum with 70% training data, indicating optimal generalization occurs when the models are trained with 70% of the available data.

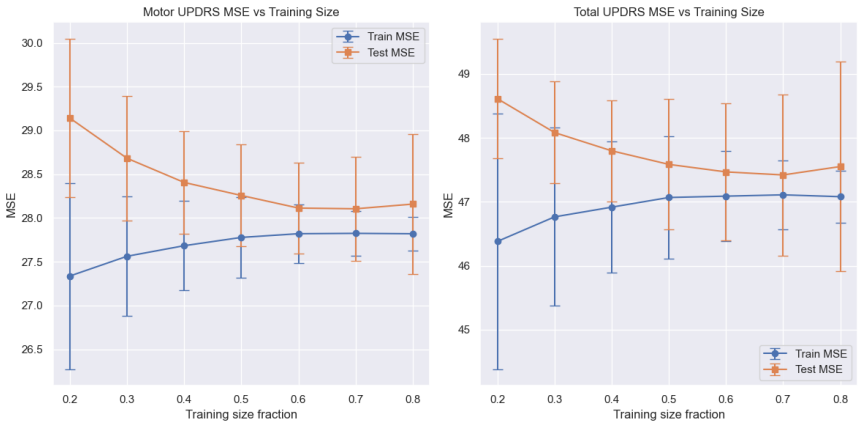

Fig 1. Mean Squared Error vs. Training Size Fraction

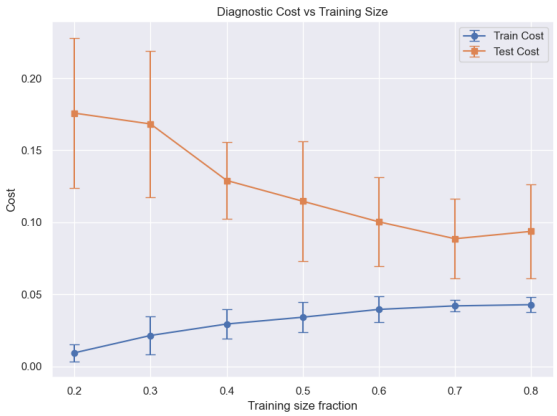

Fig 2. Cost vs. Training Size Fraction

4.4 Effect of Increasing Mini-batch Sizes of Training Data on Performance

We evaluated mini-batch SGD on our models with increasing mini-batch sizes (powers of two from 4 until 512 and full batch) revealing a distinct, dataset-dependent effect. For the Parkinson dataset using linear regression (Figure 3), performance generally improved (with the exception of batch size 8), with full-batch gradient descent achieving the optimal test MSE (27.997 for Motor_UPDRS, 46.544 for Total_UPDRS); however, large batch sizes of 64 or larger yielded nearly identical results, indicating a performance plateau. In contrast, the Breast Cancer dataset (Figure 4) showed robustness to batch size, with the test cost remaining consistently low across values. Although batch size 4 achieved the numerically lowest test cost of 0.0887, the practical difference was negligible, confirming the stability of its optimization landscape and the hyperparameter's minimal impact on final model performance.

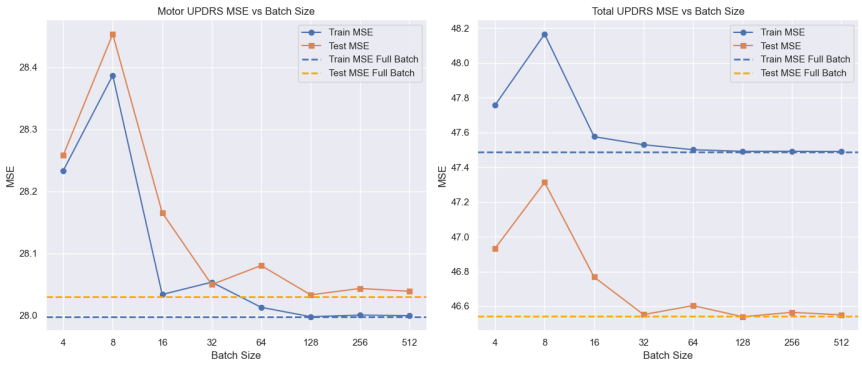

Fig 3. MSE vs. Batch Size

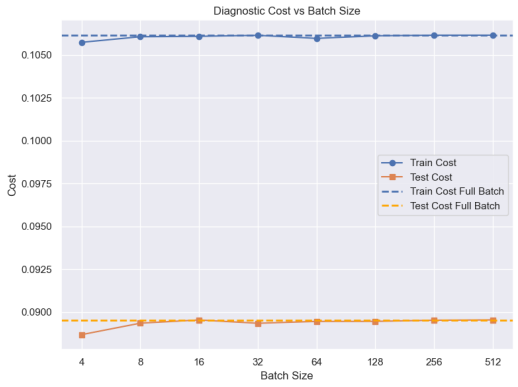

Fig 4. Cost vs. Batch Size

4.5 Effect of Learning Rate on Performance

We tested learning rates {0.001, 0.005, 0.01, 0.05, 0.1} using full-batch gradient descent with epsilon = 1e-20 and max_iter = 1000. For the Parkinson dataset (Figure 5), a learning rate of 0.1 is optimal for linear regression, achieving the lowest test MSE (27.99 for Motor UPDRS). Similarly, for the breast cancer dataset (Figure 6), a learning rate of 0.1 yielded the best performance for logistic regression, reaching the lowest test cost (0.056). In both cases, the largest learning rate provided optimal performance with fast convergence and no overfitting, while the smallest learning rate (0.001) converged too slowly and failed to reach a steady state within the iteration limit.

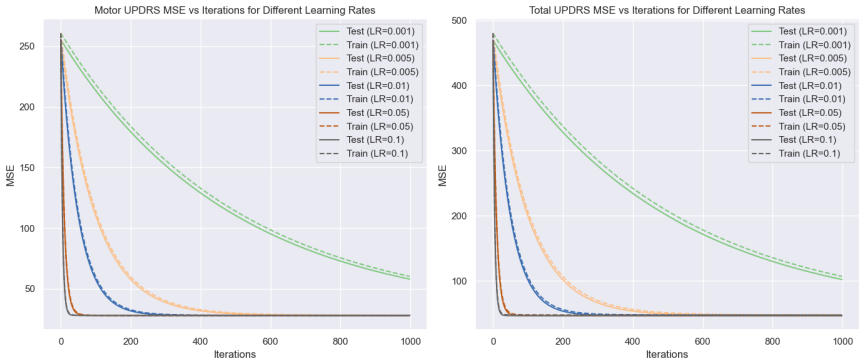

Fig 5. MSE vs. Iterations w.r.t. learning rate

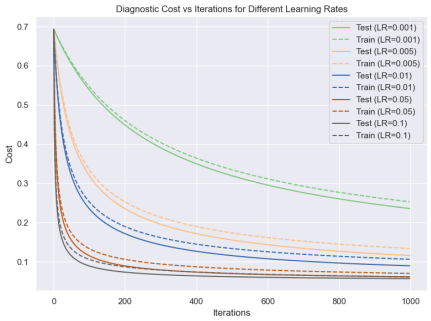

Fig 6. Cost vs. Iterations w.r.t. learning rate

4.6 Analytical linear regression vs. Mini-batch SGD linear regression

We compared closed-form linear regression solution mini-batch SGD across batch sizes {8, 16, 32, 64, 128, 256, 512} using an 80/20 train/test split (Figure 7). For Motor UPDRS, the analytical linear regression achieved a lower MSE. However, for Total UPDRS, a different pattern emerged: the analytical solution was outperformed on the test set by mini-batch SGD for batch sizes ≥ 32 (MSE Test Analytical is 46.653 while MSE for Total_UPDRS is 46.552 at batch size 256). This indicates that for the Total UPDRS target, the analytical solution is more likely to overfit, and the iterative regularizing effect of SGD with larger batch sizes provides a better generalizing model.

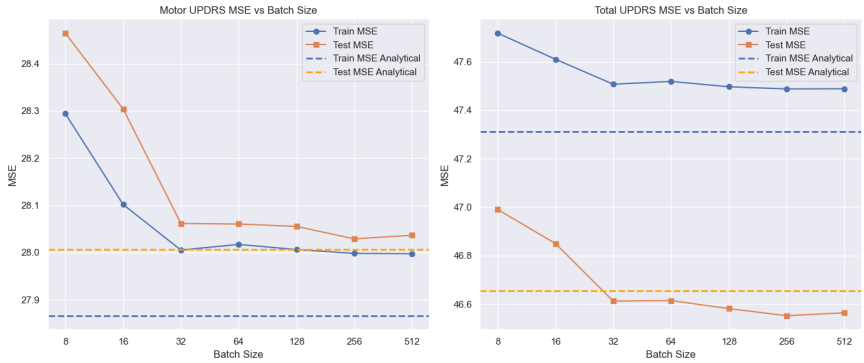

Fig 7. MSE vs. Batch Size compared to Analytical MSE

4.7 Combinatorial try-outs of features

We hypothetized that using the full set of features may not be optimal when training a model on a dataset, and that using smaller subsets might improve model performance. Considering that both the *Parkinson's Telemonitoring* and the *Breast Cancer Wisconsin* datasets are small, we decided to try out all subsets of feature using a combinatorial algorithm. This is obviously very computational intensive, and for larger datasets with more features/examples, we would've opted for a smarter selection of features.

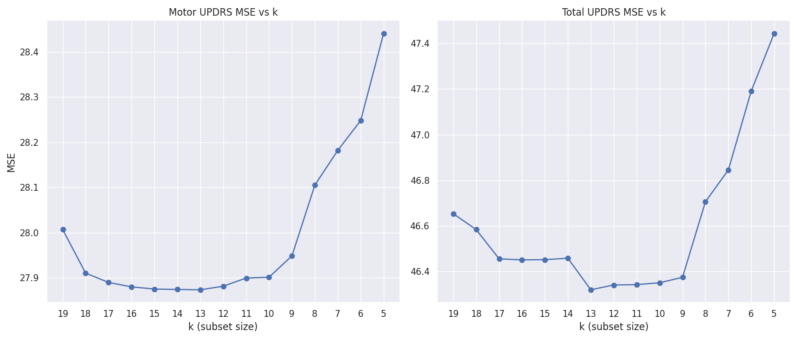

Fig 8. MSE vs. Num. of features

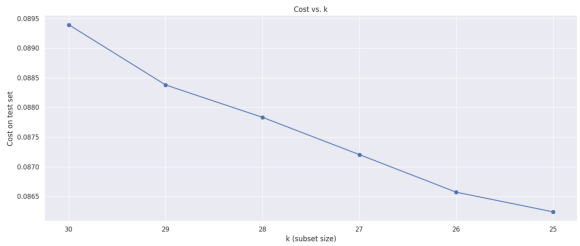

Fig 9. Cost vs. Num. of features

This experiment revealed that for linear regession with the *Parkinson's telemonitoring* algorithm, a subset of 13 features was optimal for both Total UDPRS and Motor UDPRS (Figure 8), although it was not the same subset for both targets. As for logistic regression with *Breast Cancer Wisconsin*, a subset of 25 features performed the best (Figure 9). An even smaller subset would've presumably been optimal, but we didn't have the necessary resources to compute it.

5 Discussion and Conclusion

This project showed what basic linear and logistic regression can do on two real-life medical datasets when built from scratch. On the *Parkinson's Telemonitoring* dataset, our linear regression model made reasonable predictions: test errors were about MSE ≈ 28 for *motor_UPDRS* and ≈ 46.7 for *total_UPDRS*.

We also saw that mini-batch SGD could match the analytical solution, and for *total_UPDRS*, even beat it. This can possibly be explained by the small, imperfect GD updates and by the ability to stop early, which can both prevent "overfitting" of the model. Using more training data helped overall (with the best split being 70/30), and small to medium batch size (16-256) performed as well or better than using the full batch.

On the *Wisconsin Breast Cancer Diagnostic* dataset, logistic regression reached high test accuracy with a cost as low at 0.056. This is encouraging because logistic regression is simple and easy to explain, yet already strong for this real-life task.

Looking ahead, there are several improvements to try:

- For the Parkinson's data, split by patient (so the same person never appears in both train and test). This better reflects real use and avoids optimistic results.
- Try non-linear models such as decision trees, which could learn patterns that a straight line can't.
- Improve experiment 4.7 by 1. parallelizing feature-subset evaluation across CPU cores to reduce runtime, and 2. pruning the search space with informed subset selection.

Overall, our results showed that well-tuned basic models are a solid starting point. Small but careful changes to the preprocessing could potentially make them more reliable for real-life uses.

6 Statement of Contributions

Nicholas Milin contributed to the implementation and analysis of linear regression and mini-batch stochastic gradient algorithms. Diana Covaci contributed to the implementation and analysis of logistic regression. Nicholas Milin, Diana Covaci, and Carl-Elliott wrote the report. Carl-Elliott implemented experiment 4.7 and converted the report to LaTeX.

References

[1] Abien Fred Agarap. On breast cancer detection: An application of machine learning algorithms on the wisconsin diagnostic dataset. *arXiv preprint arXiv:1711.07831*, 2017.

[2] I. Bavli, J. Smith, and A. Lee. Ethical concerns around privacy and data security in AI health monitoring for Parkinson's disease: Insights from patients, family members, and healthcare professionals. *Journal of Medical Ethics*, 51(2):123–130, 2025.

[3] Ömer Eskidere, Figen Ertaş, and Cemal Hanilçi. A comparison of regression methods for remote tracking of Parkinson's disease progression. *Expert Systems with Applications*, 39(17):15982–15990, 2012.

[4] P. A. Napitupulu. Ethical dilemmas in the use of artificial intelligence in breast cancer diagnosis and treatment: Addressing issues of bias, transferability, and patient trust in breast cancer AI. *West Science Law and Human Rights*, 1(04):256–260, 2025.

[5] Ahmed F. Seddik and Doaa M. Shawky. Logistic regression model for breast cancer automatic diagnosis. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 642–646. IEEE, 2015.

[6] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, 1993.

[7] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, 2009.