Powerstats

Predicting drug use in powerlifting based on performance

CS4365/6365 Introduction to Enterprise Computing (Su'2025)

Brandon Manville

Introduction

- Powerlifting - Squat, Bench, and Deadlift
    - 3 attempts at each lift
    - Score is determined by the sum of the best attempt from each lift
- Project Goal: Predicting drug use in powerlifting from sports performance alone
- Deliverables:
    - Models capable of predicting drug use from performance metrics alone
        - Needs to be better than randomly guessing
    - Data analysis and machine learning pipeline
    - CLI/GUI to access models
    - LLM interface to analyze powerlifting performance data

Related Work

- No work directly done on powerlifting
- Ryoo et al. (2024) 53.8% prediction rate among the weightlifters
    - XGBoost, Multilayer Perceptron (MP), Ensemble (XGBoost + MP)
- Hopker et al. (2024) present evidence that competitive performance alone can discriminate between doped and clean athletes with significant accuracy in a Bayesian framework

Significance

- Drug use is rampant and standards are low
- USA Powerlifting drug tests <=10% of athletes randomly
    - Sometimes podium finishes are tested
- Ayotte et al (2013) report on World Anti Doping Agency (WADA): 2,790 adverse analytical results from 258,267 tests analyzed
    - That's only 1.08%
        - a significant portion of those results were for cannabis only
- 14%–39% is likely the prevalence of intentional doping in elite sports (de Hon, 2015)
- Huge gap - something needs to change!

Datasets

- Open Powerlifting (training/validation)
  - https://www.openpowerlifting.org/
  - In depth description: https://openpowerlifting.gitlab.io/opl-csv/bulk-csv-docs.html
  - We care about SBD, total, sex, age, bodyweight, and if the competition was tested/untested
  - Use sequences of competition results as input and output if the meet was tested/untested in training
- USAPL Drug Testing Database (testing)
  - https://www.usapowerlifting.com/drug-testing/
  - Tuples of names, dates, and drug testing results
  - Match against Open Powerlifting to replace the meet tested/untested label with the drug testing results

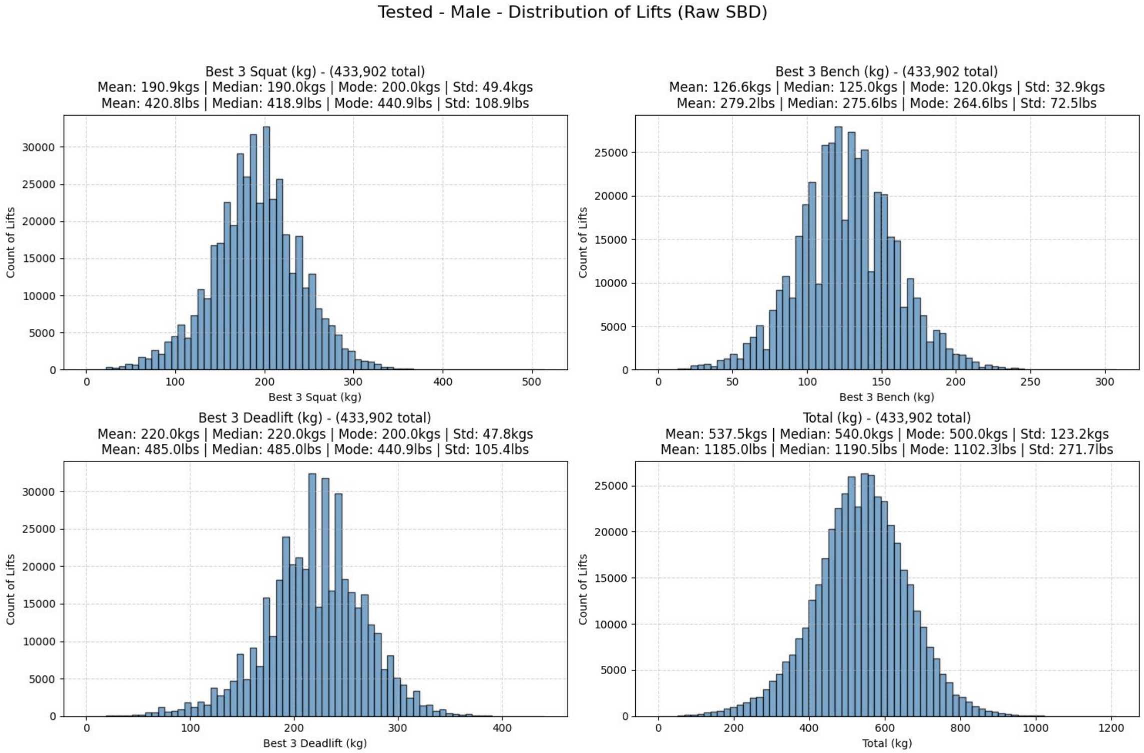
Data Labeling

- Goal: Predict drug use
- Problem: *Very* limited data labeled for drug use
  - Competitions don’t report this
  - USAPL drug use dataset has roughly 200 positive drug test results from 2018-2025 and only 111 of those could be cross-referenced with Open Powerlifting

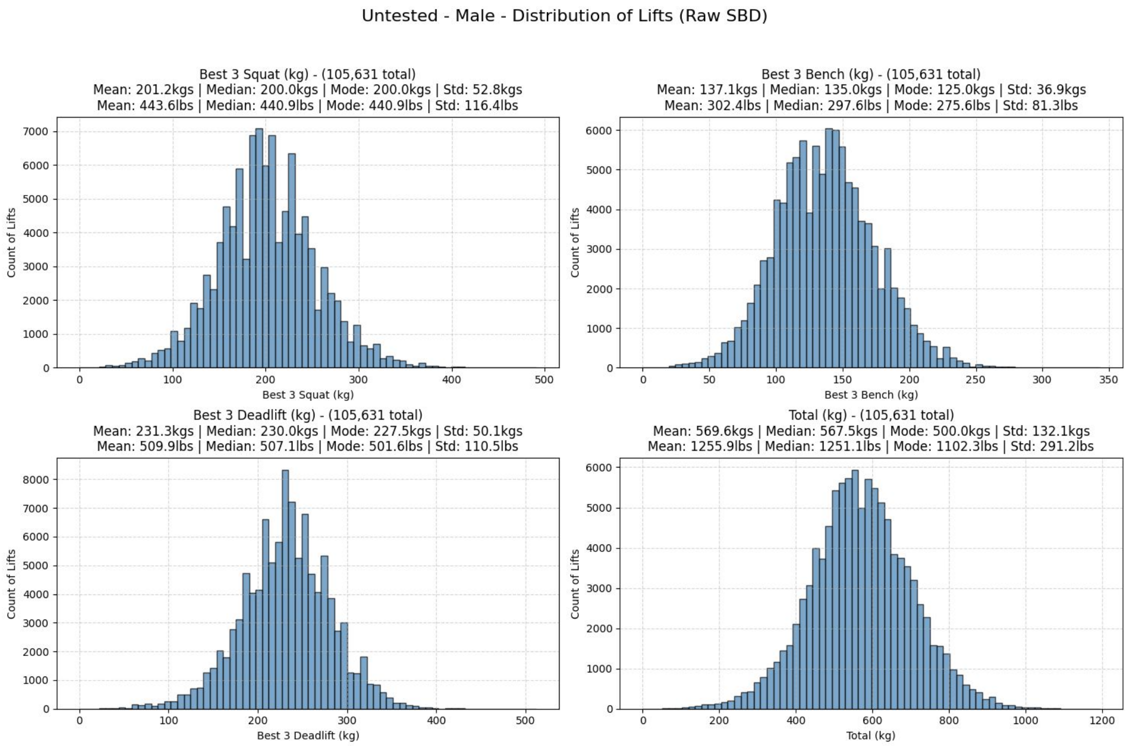
Distributions of Lifts

Let's look at some distributions of lifts from male and female lifters from tested and untested competitions.

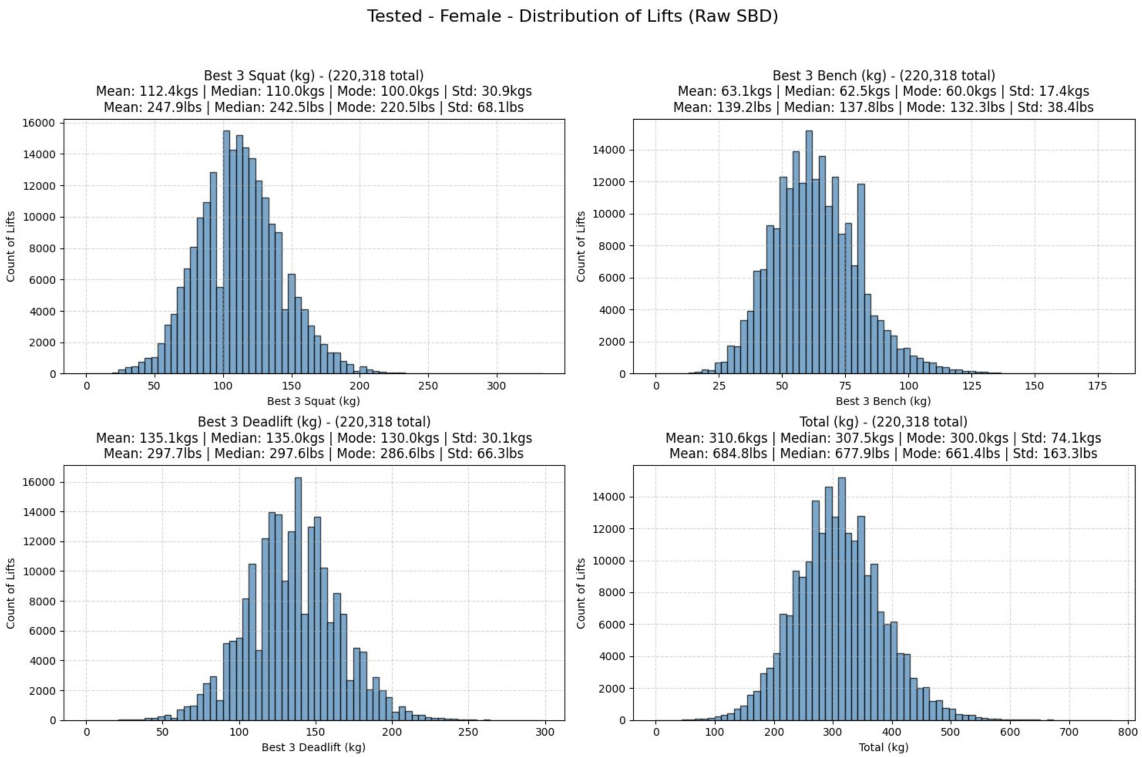
Distribution of Lifts (Male - Tested)



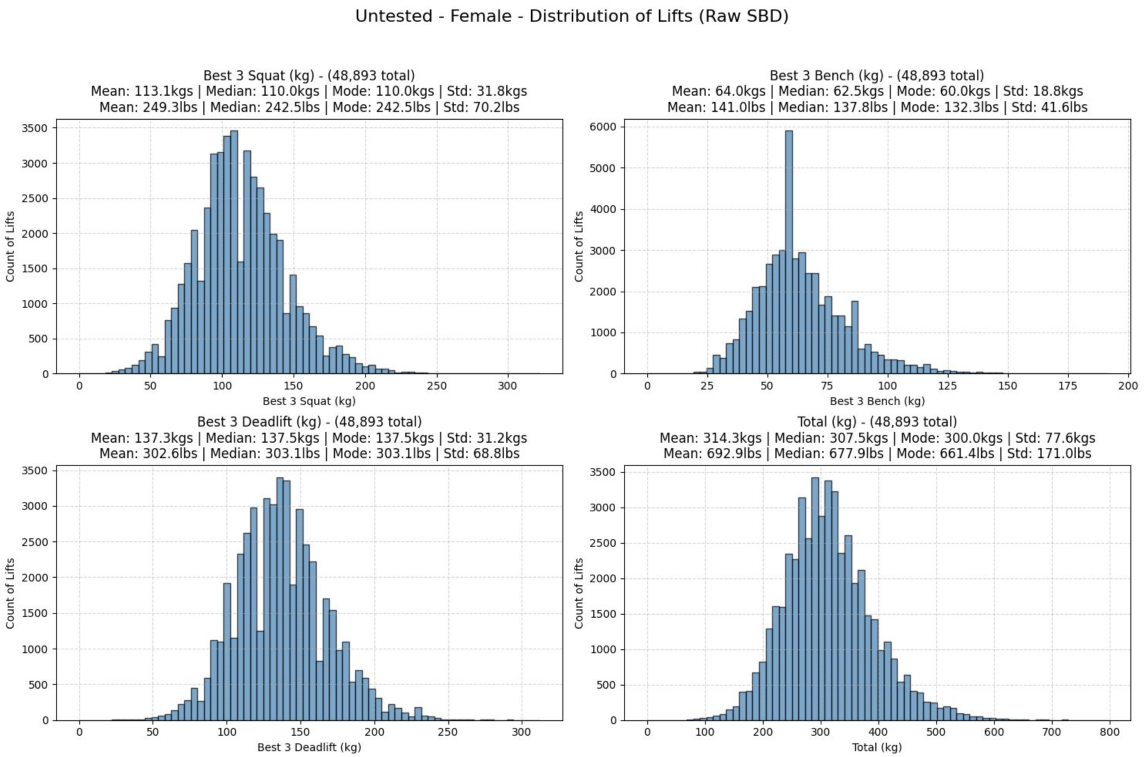
Distribution of Lifts (Male - Untested)



Distribution of Lifts (Female - Tested)



Distribution of Lifts (Female - Untested)



A Shift of Statistics

- Male Tested Average Total: 537.5kg
- Male Untested Average Total: 569.6kg
- Female Tested Average Total: 310.6kg
- Female Untested Average Total: 314.3kg

-> Untested lifters have an advantaged (PEDs). Can we learn from this?

Data Labeling (Revisited)

- Goal: Predict drug use
- Problem: *Very* limited data labeled for drug use
- **Solution: Train to predict tested/untested meets as a proxy for drug use**

Data Input/Cleaning

- Predict drug use from performance metrics
  - Squat, Bench, Deadlift, Total, Bodyweight, Age, Sex
- Filter all invalid results from database
- Cross-reference USAPL Drug Testing database consisting of names and drug testing results against Open Powerlifting database to create test data
- Input to the models will be the history of all results in order up to a meet
  - I.e. X=[Performance at meet 1, Performance at meet 2, …, Performance at meet n]
  - Y= Using drugs at meet n
- Training label will be if the last meet in the sequence was tested (0) or untested (1)
- Testing label will be if the athlete tested negative for PEDs (0) or positive (1) at the last meet in the sequence

What’s Good at Analyzing Data over Time?

- RNNs!
- 3 models trained and tested:
  - Classic RNN
  - LSTM
  - Bi-directional LSTM
- But also, LLMs aren’t bad either on time data
  - Particularly, the transformer model is really good at this
  - LLMs are good at text data so results might not be good here

Training the Models

- After cleaning the data, 308,464 training sequences created
    - Percent drug tested: 50.00%. Percent not drug tested: 50.00%.
- After cross-referencing with Open Powerlifting, created 111 true drug-use labels and 111 true drug-free labels
- Models were trained on 80/20 train/validation split of training data
    - Validation set was used for hyperparameter grid search
- Models were tested on the 222 confirmed labels

Grid Search for LSTM First

```
# Hyperparameter grids
# for the initial LSTM using SGD optimizer and
param_grid: dict[str, list[float | int]] = {
    "hidden_size": [32, 64, 128, 200],
    "num_layers": [1, 2, 3],
    "dropout": [0.0, 0.1, 0.3],
    "lr": [0.0001, 0.0005, 0.001, 0.01],
}
# the best models on the full dataset (80/20 split) are (that I manually selected by F1 score > ~0.6 and accuracy > 0.56):
#
# hidden_size,num_layers,dropout,lr,f1,accuracy,precision,recall
# 200,1,0.0,0.001,0.6174961593089655,0.5721973675679181,0.5575971731448763,0.6918130744000259
# 128,1,0.0,0.001,0.618791449634247,0.565810802048888,0.5507562131076938,0.706004611437664
# 200,2,0.1,0.001,0.6037984507475362,0.5614180120599105,0.5498159901861432,0.6695352839931153
# 200,2,0.0,0.01,0.5969201413216857,0.5617097840886986,0.5517250881834215,0.6501802357678693
```

Grid Search Again for all Model (Narrowed)

```
param_grid: dict[str, list[float | int]] = {
    "hidden_size": [128, 200, 256, 384],
    "num_layers": [1, 2],
    "dropout": [0.0, 0.1],
    "lr": [0.001, 0.0005],
}
```

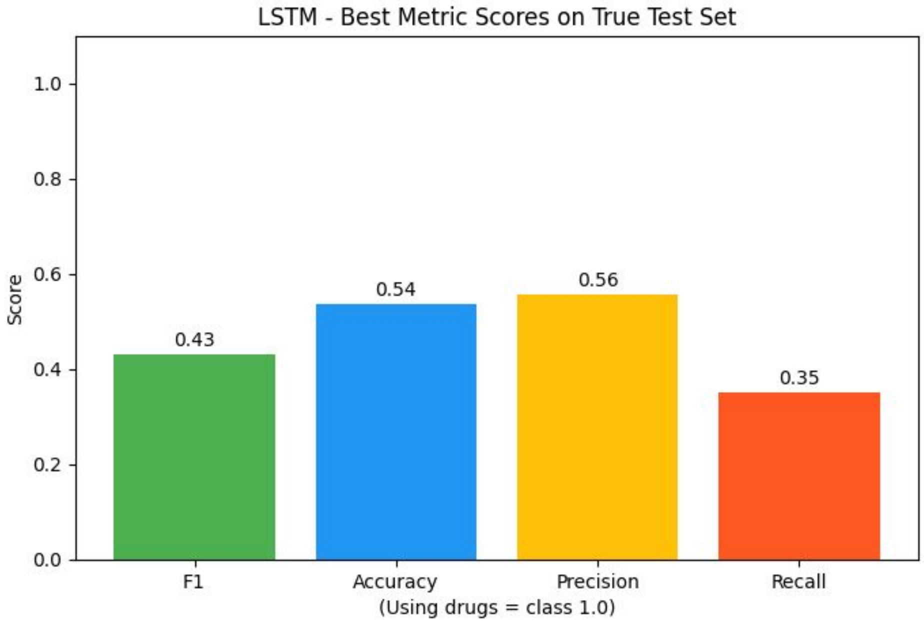
Results on the Validation Sets

Best Hyperparameters for Bidirectional_LSTM: Hidden Size: 128 Num Layers: 2 Dropout: 0.1 Learning Rate: 0.0005 Final Evaluation on Validation Set: Accuracy: 0.6003 Precision: 0.5763 Recall: 0.7527 F1: 0.6528	Best Hyperparameters for LSTM: Hidden Size: 384 Num Layers: 2 Dropout: 0.0 Learning Rate: 0.0005 Final Evaluation on Validation Set: Accuracy: 0.5981 Precision: 0.6229 Recall: 0.4943 F1: 0.5512	Best Hyperparameters for RNN: Hidden Size: 200 Num Layers: 2 Dropout: 0.1 Learning Rate: 0.0005 Final Evaluation on Validation Set: Accuracy: 0.5855 Precision: 0.5846 Recall: 0.5868 F1: 0.5857

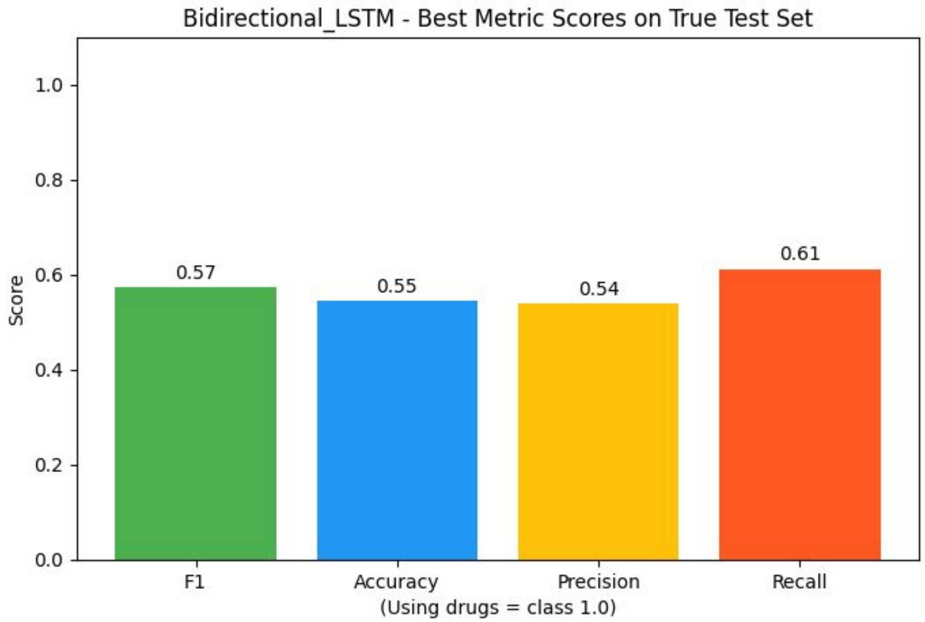
Now Time to Test!

Drum roll...

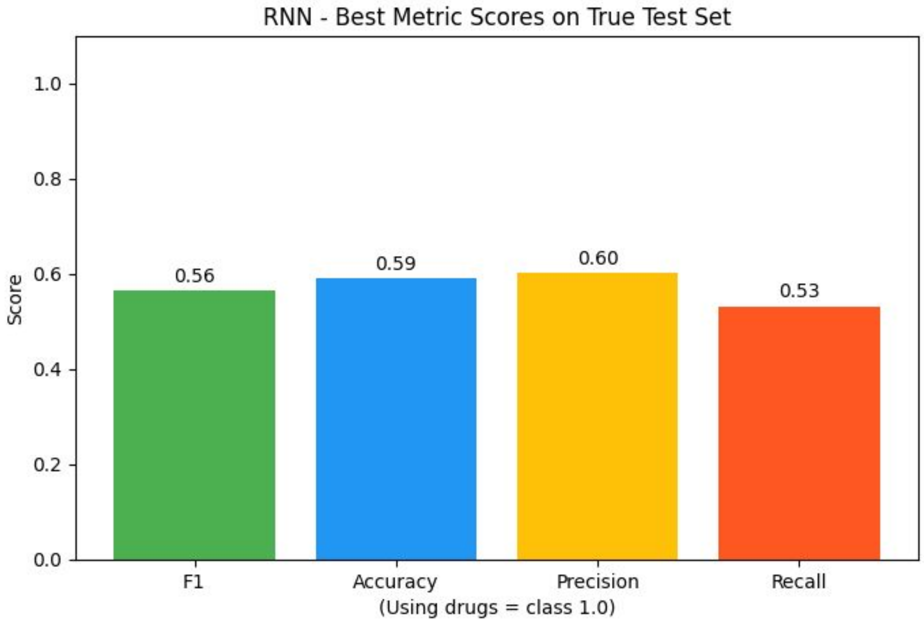
Results - LSTM



Results - Bidirectional LSTM

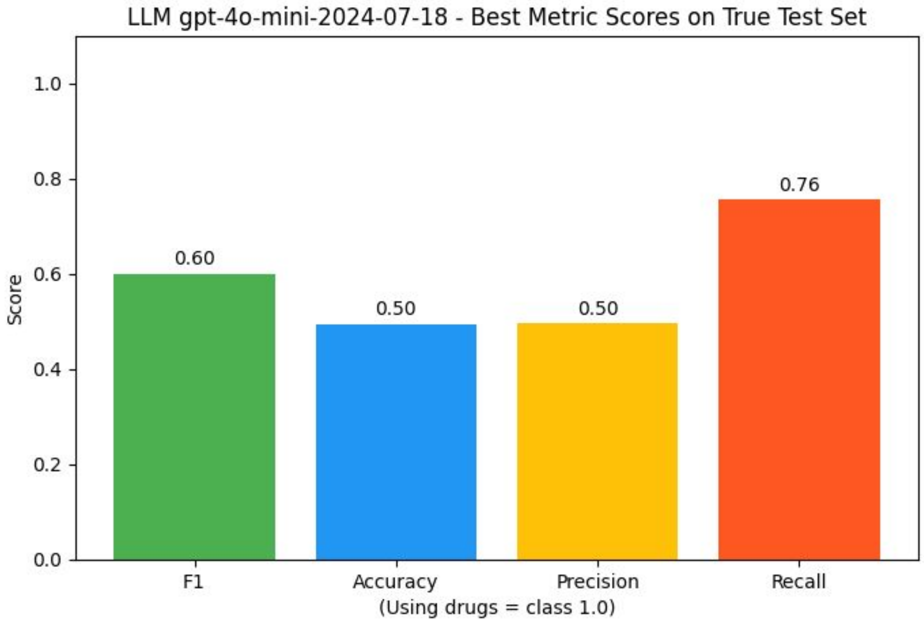


Results - RNN



RNN - Best Metric Scores on True Test Set
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.56
0.59
0.60
0.53
F1
Accuracy
Precision
Recall
(Using drugs = class 1.0)

Results - LLM (GPT 4o-mini)



Rankings (by accuracy)

1. RNN - 59% Accuracy and 0.56 F1 Score **(winner)**
2. Bidirectional LSTM - 55% Accuracy and 0.57 F1 Score
3. LSTM - 54% Accuracy and 0.43 F1 Score
4. LLM (GPT 4o-mini) - 50% Accuracy and 0.60 F1 Score
    - (and it took 25 minutes, the others took <1 minute, and worst of all costs money to use)

How Does It Compare?

- Recall Ryoo et al. (2024) 53.8% prediction rate among the weightlifters
- Random guessing is 50%
  - The current standard in powerlifting (and most other spots)
- We succeed in creating a better method
  - But it’s certainly not the the bee's knees
  - Best model was 59% accuracy on true data
  - Models seem to be unsure (predictions are often around 0.4-06 range)

Conclusions

- This project has shown that there is potential for predicting drug use in powerlifting on performance metrics alone
- Models can assist referees when selecting candidates for drug testing
- The average LLM still isn't good with numbers
- Regardless of the results of this project, there needs to be a change in drug testing standards!

Future Work

- Future work should be dedicated to collecting more data with true drug testing labels
  - I believe this is by far the limiting factor
  - Too much variance in lifters who compete in untested competitions (a lot of them are just casual lifters!)
- Data engineering
- Incorporating biological markers
- More complex models
- Fine tune LLMs if more data becomes available

References

- Hyunji Ryoo et al. “Identification of doping suspicions through artificial intelligence-powered analysis on athlete’s performance passport in female weightlifting”. In: Frontiers in Physiology Volume 15 - 2024 (2024). ISSN: 1664-042X. DOI: 10.3389/fphys.2024.1344340. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2024.1344340.
- James G. Hopker et al. “Competitive performance as a discriminator of doping status in elite athletes”. In: Drug Testing and Analysis 16.5 (2024), pp. 473–481. DOI: https://doi.org/10.1002/dta.3563. eprint: https://analyticalsciencejournals.onlinelibrary . wiley . com / doi / pdf / 10 . 1002 / dta . 3563. URL: https : / /analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/dta.3563.
- C Ayotte et al. “Report to WADA Executive Committe on Lack of Effectiveness of Testing Programs”. In: Montreal: WADA (2013).
- Olivier de Hon, Harm Kuipers, and Maarten van Bottenburg. “Prevalence of Doping Use in Elite Sports: A Review of Numbers and Methods”. In: Sports Medicine 45.1 (2015), pp. 57–69. DOI: 10 . 1007 / s40279 - 014 - 0247 - x. URL: https ://link.springer.com/article/10.1007/s40279-014-0247-x.

GitHub

- https://github.com/bmanville3/powerstats

Demo of CLI/GUI

- The software is self-explanatory for the most part but here is a short demo