CS 4365 Final Project: Powerstats

An analytical approach to predicting drug use in powerlifting.

Brandon Manville

July 24, 2025

Contents

1 Introduction

The sport of powerlifting has seen significant growth in recent years, with competitions and athlete participation expanding across various federations. One major concern that continues to plague the sport is the use of performance-enhancing drugs (PEDs), which compromise fair competition and athlete health. Doping detection is a particularly difficult task as not all atheletes are drug tested, and, if they are tested and are doping, it is not necessary true the test will come back positive. Therefore, this project aims to develop a method for predicting potential drug use in powerlifting athletes by analyzing performance data, particularly using machine learning techniques for drug use prediction and leveraging LLMs to summarize performance data. This project explores the feasibility and performance of models that can help identify potentially anomalous athlete profiles based on publicly available competition data.

This project aim to produce the following deliverables:

1. A model capable of predicting drug use from performance metrics alone that is better than randomly guessing.

2. A data analysis and machine learning pipeline capable of training and testing models on powerlifting performance data.

3. A CLI/GUI to access the models and perform anlyses on powerlifting performance data.

4. An LLM interface to analyze powerlifting performance data.

2 Motivation

The use of performance-enhancing drugs remains one of the most persistent and complex challenges in competitive sports. Traditional anti-doping measures, such as randomized drug testing or testing based on podium finishes, are often inconsistent and susceptible to oversight. This is evident in the sport of raw, tested powerlifting, where detection of drug use is both difficult and underdeveloped. Unlike many sports that rely on biological passports [1] or detailed lab testing regimes, powerlifting federations-such as the United States of America Powerlifting (USAPL)-typically test only a small percentage of competitors (typically $\leq 10\%$), often chosen at random or based on exceptional performance. This leaves a considerable blind spot for potential PED use among athletes who do not stand out dramatically or simply fall outside of the arbitrary performance threshold the judges have decided on the day.

This project proposes a novel, data-driven approach to detect potential drug use in raw powerlifting based solely on performance metrics. By leveraging large-scale competition data from the Open Powerlifting dataset [2], which includes over 3.6 million lift entries from more than 556,000 athletes, the goal is to develop machine learning methods that can identify anomalous performance trends potentially indicative of PED use. However, this task is particularly challenging due to the lack of ground truth labels-there is little publicly available data on who has used PEDs, who was tested, or who failed a drug test. Furthermore, athlete performance is influenced heavily by individual differences, especially genetics, which adds further noise to any attempt at prediction.

3 Related Work

Previous research on doping detection has largely focused on olympic sports, like cycling, weightlifing, or shotput, where longitudinal data on blood profiles enables the use of the Athlete Biological Passport (ABP) [1]. In strength sports such as powerlifting or strongman, such approaches are rare due to the lack of systematic testing and data. Furthermore, at present there have been no papers published on predicting drug use from powerlifting performance. Hence, we will focus on work from other sports in this section.

Recent literature in other fields have begun to emphasize the integration of performance-based data into anti-doping analytics. Hopker et al. (2020) propose a paradigm shift towards performance profiling as a complementary intelligence-led strategy, enabling the detection of atypical results suggestive of performance-enhancing substance use [3]. Similarly, Iljukov and Schumacher (2017) argue for the value of longitudinal performance profiling in identifying outliers in athletic performance trajectories [4].

The Athlete's Performance Passport (APP) has emerged as an extension of the ABP, focusing specifically on competitive results. Applications of the APP have been explored in various domains, such as in the work of Iljukov et al. (2018), who detail its use in a real-world doping investigation case [5]. Faiss et al. (2019) further advocate for modeling performance as a tool for doping control, asserting that analytics can help differentiate between natural and artificial progression [6].

Bayesian modeling has gained traction as a statistical method suited for such detection. Montagna and Hopker (2018) demonstrate a Bayesian framework that quantifies the probability of performance enhancement for shotput athletes being due to doping, offering a probabilistic alternative to rigid threshold models [7]. In a more recent study, Hopker et al. (2024) present evidence that competitive performance alone can discriminate between doped and clean athletes with significant accuracy in a Bayesian framework [8].

The incorporation of artificial intelligence (AI) in doping detection marks another frontier. Ryoo et al. (2024) employed XGBoost, Multilayer Perceptron (MLP), and an Ensemble model, which integrates XGBoost and MLP, models to analyze female weightlifters' performance patterns, revealing suspicious anomalies aligned with known doping behaviors, which achieved a "53.8% prediction rate among the weightlifters sanctioned in the 2008, 2012, and 2016 Olympics" [9]. Complementarily, Ryoo et al. (2022) reinforce the importance of performance data in weightlifting, demonstrating that longitudinal analysis can detect doping more efficiently than traditional methods [10].

Despite these advancements, criticisms persist regarding the efficacy of global testing programs. Quantifying the true prevalence of doping remains difficult due to underreporting and methodological gaps. de Hon et al. (2015) provide a comprehensive review of prevalence estimation techniques, underscoring how indirect methods, such as survey-based assessments and longitudinal performance analysis, consistently reveal higher doping rates than testing data alone suggest. In this paper, it is estimate that "14%–39% is likely to be a more accurate reflection of the prevalence of intentional doping in elite sports" [11].

A report to the World Anti-Doping Agency (WADA) expressed concern over the lack of effectiveness in conventional testing approaches, urging the adoption of more intelligent systems [12]. From the report: "The latest available laboratory statistics (for 2010) indicate a mere 2,790 adverse analytical results were returned from more than 258,267 tests analyzed; a meager 1.08%. In addition, a significant portion of those results were for cannabis only" [12]. Empirical studies echo this concern; Aguilar-Navarro et al. (2020) found that many sports exhibit disproportionately low Adverse Analytical Findings (AAFs) despite doping suspicions, pointing to systematic underdetection [13].

The 14%–39% predicted by de Hon et al. is staggeringly higher than the 1.08% of AFFs report with many cases being for cannabis. Therefore, it is imperative that a smarter system of drug testing be implemented.

4 Datasets

4.1 Open Powerlifting Database

The primary dataset used in this project is the Open Powerlifting dataset [2], a comprehensive, open-source database containing competition results from powerlifting meets worldwide. For detailed documentation, readers are referred to the following resources:
https://gitlab.com/openpowerlifting/opl-data/blob/main/docs/data-readme.md.

Each row in this dataset corresponds to a single lifter's performance at a specific meet. From each entry, we extracted the following attributes for model input:

1. Best successful attempt in the squat, bench press, and deadlift (in kilograms),
2. Total weight lifted (in kilograms),
3. Bodyweight (in kilograms),
4. Age (in years),
5. Sex (binary encoded: 1 for male, 0 for female),
6. Meet type (binary encoded: 0 for drug-tested, 1 for untested).

Other fields such as competition date, Wilks score, and DOTS score were retained for exploratory analysis but not used during model training.

4.2 USAPL Drug Testing Database

In addition to this dataset, we incorporated the USA Powerlifting (USAPL) drug testing records, which are publicly available and include detailed information about all administered drug tests from 2004 onward [14]. For the purposes of this project, we used records from 2018 to 2025, which provided labeled instances of lifters who tested positive or negative for banned substances. From this dataset, we extracted lifters' names, testing dates, and outcomes, and cross-referenced these with entries in the Open Powerlifting dataset to construct labeled test cases.

It is important to note that both datasets are periodically updated. Therefore, exact replication of our results may yield minor discrepancies depending on the version of the dataset used.

5 Methodology

5.1 Data Cleaning

After downloading the Open Powerlifting dataset, we filtered the data to include only valid entries corresponding to **raw, full-power** competition results. Raw lifting refers to performances conducted without the use of assistive equipment such as squat suits or bench shirts, but allowing accessories like belts, knee sleeves, and wrist wraps. Full-power meets require lifters to complete all three lifts: squat, bench press, and deadlift.

We further restricted the dataset to entries with non-null and physically plausible values for all required fields. The resulting cleaned dataset was imported into a SQL database to facilitate efficient querying and aggregation.

The USAPL drug testing records were parsed into a structured format containing tuples of the form (name, test date, test result). These names were then matched against the Open Powerlifting dataset. For each matched lifter, we extracted all performance records up to the date of the drug test to construct labeled sequences for model evaluation.

5.2 Data Input - Creating Data Sequences

For each lifter in the dataset, we constructed time-ordered sequences of meet results. These sequences served as input examples for the models. Consider a lifter with the following chronological competition history:

$$R_0, R_1, ..., R_n$$

where each R_i represents a performance at a single meet. From this history, we generated a set of progressively growing sequences and their corresponding labels (based on the meet type):

$$([R_0], y_0),\ ([R_0, R_1], y_1),\ \dots\ ([R_0, R_1, \dots, R_n], y_n)$$

where each $y_i \in \{0, 1\}$ indicates whether the final competition in the sequence was conducted at a tested (0) or untested (1) meet.

5.3 Data Labeling and Model Training

A major limitation of the Open Powerlifting dataset is the absence of individual-level drug testing labels. However, it does indicate whether a meet was officially a drug tested meet. While this is not a perfect proxy-since not all participants in a drug-tested meet are necessarily tested-it provides a valuable signal that can be leveraged for supervised learning.

To support the validity of this proxy, we compared the distributions of performance totals across tested and untested meets, stratified by sex. These distributions (visualized in Figures 1-4) revealed notable differences in average total weights lifted:

Male (Tested): 537.5 kg **Male (Untested):** 569.6 kg
Female (Tested): 310.6 kg **Female (Untested):** 314.3 kg

These discrepancies suggest that untested competitions are associated with higher performance outputs, particularly among male lifters. One plausible explanation is that some degree of performance enhancement via drug use occurs more frequently in untested meets. Thus, we posit that a model trained to predict whether a lifter competed in a tested or untested meet can serve as an approximate classifier for potential drug use.

It must be emphasized, however, that this is a proxy task-not a direct prediction of drug use. Many athletes may choose competition types based on factors unrelated to substance use (e.g., geographic proximity, federation affiliation, or scheduling). To partially validate this approach, we use the small subset of lifters for whom true drug testing outcomes are available (via the USAPL database) to assess the generalizability and robustness of our models.

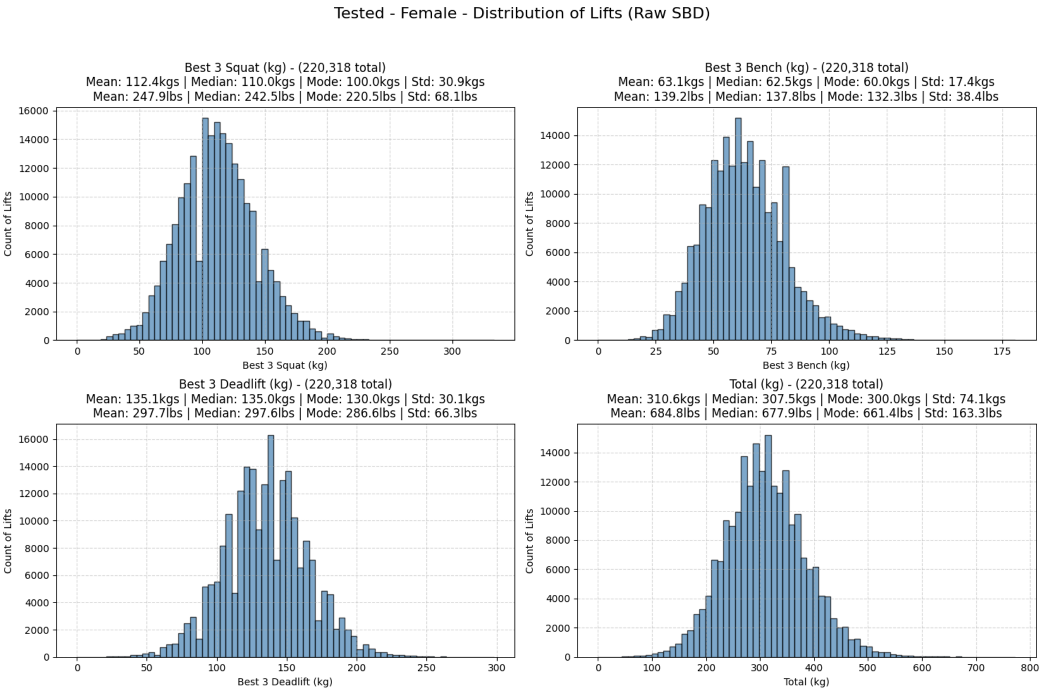


Figure 1: Female Tested Distribution of Lifts

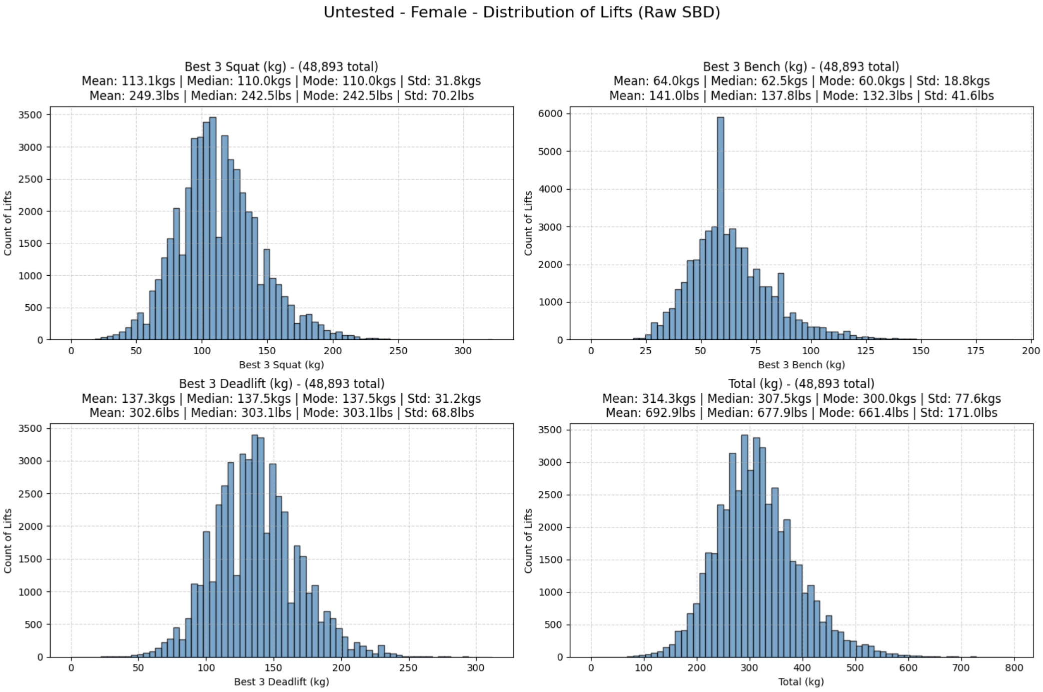


Figure 2: Female Unested Distribution of Lifts

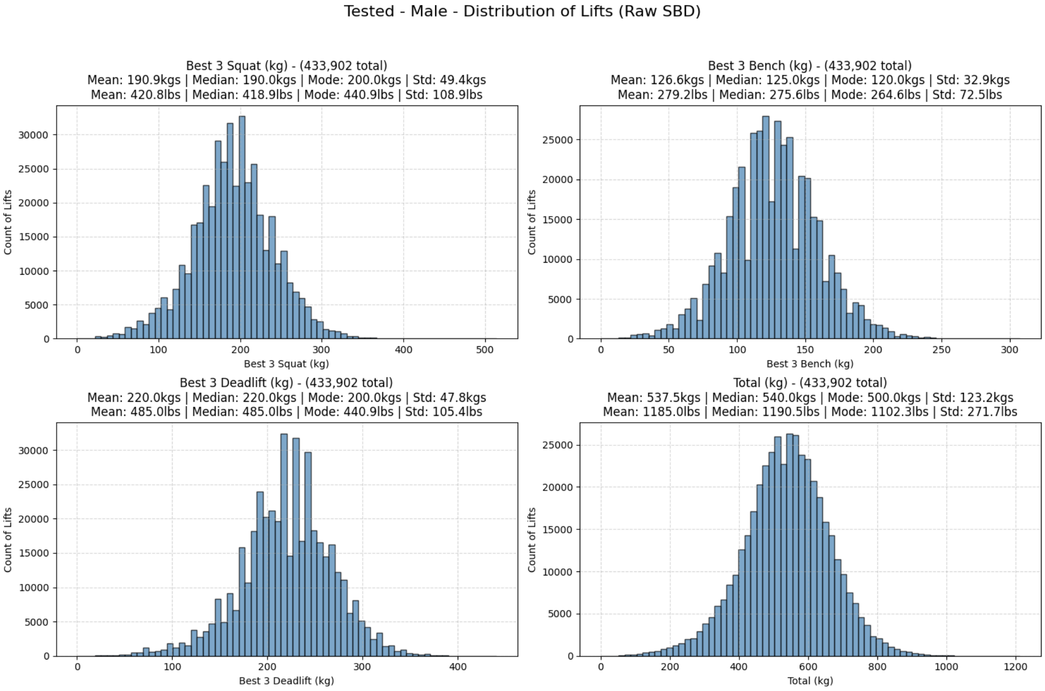


Figure 3: Male Tested Distribution of Lifts

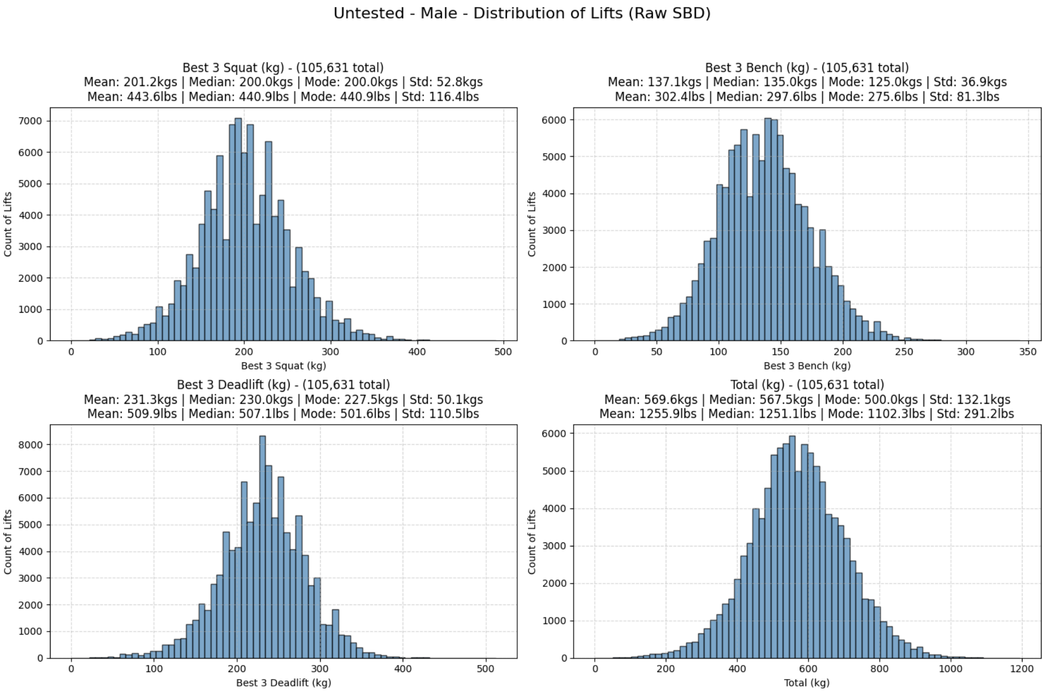


Figure 4: Male Untested Distribution of Lifts

5.4 Training-Validation-Test Split

From the cleaned and preprocessed Open Powerlifting dataset, we constructed a total of 308,464 labeled sequences individual competition results. The sequences were explicitly balanced to contain an equal proportion of tested and untested meet outcomes, yielding a 50/50 class distribution for the binary classification task.

This balanced dataset was partitioned into training and validation subsets using an 80/20 split. The training set was used to optimize model parameters, while the validation set was used for hyperparameter tuning and verifying the best models when optimizing parameters.

For evaluation on real-world drug testing outcomes, we used the USAPL drug testing database. Although the raw dataset included 199 confirmed failed tests (positive drug test results), only 111 of these could be reliably matched to lifters in the Open Powerlifting dataset. To maintain balance in the evaluation set, we selected 111 lifters who tested negative (i.e., passed drug tests), resulting in a 50/50 test set containing 222 lifters. This test set provides the only known ground-truth drug use labels and was reserved exclusively for final model evaluation.

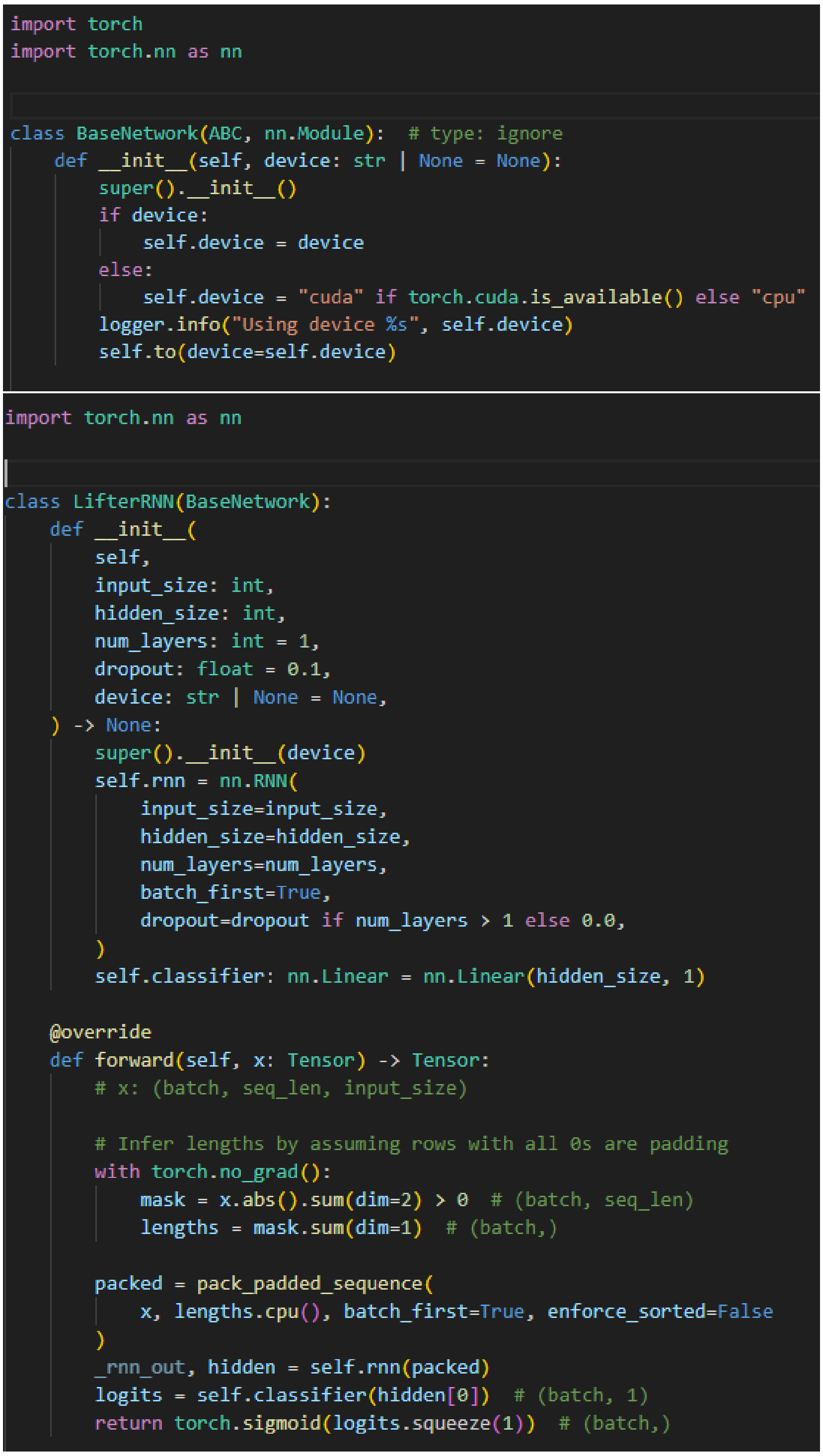
5.5 Models Used

We trained and evaluated three recurrent neural network (RRN) based architectures [15] for sequence classification:

- A standard vanilla RNN,
- A Long Short-Term Memory (LSTM) network,
- A Bi-directional LSTM (BiLSTM) network.

These models are well-suited for temporal or sequential data, such as a lifter's progression over time. PyTorch was used to implement and train all RNN-based models. A hyperparameter search was conducted (described in the following section) to optimize each model.

In addition to the RNN architectures, we also explored the feasibility of using large language models (LLMs), particularly transformer-based models, for this classification task. Although transformers are typically designed for natural language processing, they have demonstrated strong performance in time-series applications [16]. For this purpose, we utilized the OpenAI API to access pretrained transformer models without fine-tuning. We chose to use the GPT 4o-mini model as it is a fast, cost-effective with standard intelligence. Although LLMs are not ideal for this type of structured, numeric sequence data, they offer a flexible benchmark and may prove useful in future hybrid approaches.



```
import torch
import torch.nn as nn


class BaseNetwork(ABC, nn.Module):  # type: ignore
    def __init__(self, device: str | None = None):
        super().__init__()
        if device:
            self.device = device
        else:
            self.device = "cuda" if torch.cuda.is_available() else "cpu"
        logger.info("Using device %s", self.device)
        self.to(device=self.device)
```

```
import torch.nn as nn


class LifterRNN(BaseNetwork):
    def __init__(
        self,
        input_size: int,
        hidden_size: int,
        num_layers: int = 1,
        dropout: float = 0.1,
        device: str | None = None,
    ) -> None:
        super().__init__(device)
        self.rnn = nn.RNN(
            input_size=input_size,
            hidden_size=hidden_size,
            num_layers=num_layers,
            batch_first=True,
            dropout=dropout if num_layers > 1 else 0.0,
        )
        self.classifier: nn.Linear = nn.Linear(hidden_size, 1)

    @override
    def forward(self, x: Tensor) -> Tensor:
        # x: (batch, seq_len, input_size)

        # Infer lengths by assuming rows with all 0s are padding
        with torch.no_grad():
            mask = x.abs().sum(dim=2) > 0  # (batch, seq_len)
            lengths = mask.sum(dim=1)  # (batch,)

        packed = pack_padded_sequence(
            x, lengths.cpu(), batch_first=True, enforce_sorted=False
        )
        _rnn_out, hidden = self.rnn(packed)
        logits = self.classifier(hidden[0])  # (batch, 1)
        return torch.sigmoid(logits.squeeze(1))  # (batch,)
```

Figure 5: Code to Create Vanilla RNN

5.6 Hyperparameter Tuning

Initial hyperparameter tuning was performed on the LSTM model using a grid search over the following parameter space:

```
param_grid = {
    "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],
}
```

This search was executed using the training and validation datasets and required approximately 6 hours of computation. From the tuning results, we manually selected candidate configurations based on an F1 score exceeding approximately 0.6 and validation accuracy above 56%. A sample of high-performing configurations is shown below:

Hidden Size	Layers	Dropout	LR	F1	Accuracy
200	1	0.0	0.001	0.617	0.572
128	1	0.0	0.001	0.619	0.566
200	2	0.1	0.001	0.604	0.561
200	2	0.0	0.010	0.597	0.562

Based on these observations, we defined a narrower search space to be used for further model training:

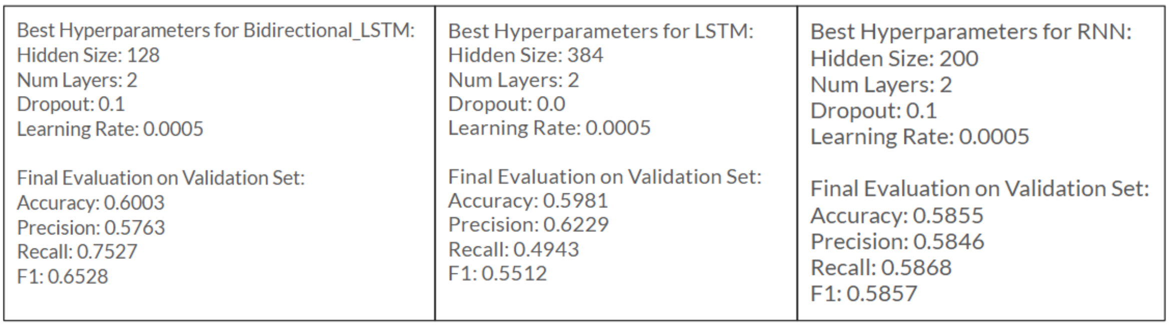
```
param_grid = {
    "hidden_size": [128, 200, 256, 384],
    "num_layers": [1, 2],
    "dropout": [0.0, 0.1],
    "lr": [0.0005, 0.001],
}
```

This refined grid reflects the following considerations:

- Larger hidden sizes ($\geq$ 128) consistently led to improved performance.
- Deeper models (3+ layers) showed diminishing returns and often overfit.
- Dropout had limited impact within the tested range, so we reduced the number of values explored.
- Learning rates near 0.001 generally performed best.

In the initial grid search, the stochastic gradient descent (SGD) optimizer was used. However, experiments with the Adam optimizer showed an immediate increase in validation accuracy of approximately 2–3%. Given the clear empirical advantage, Adam was adopted for all subsequent training runs. Although this introduces some inconsistency with the initial tuning runs, we prioritized practical performance due to computational constraints that prevented repeating the entire grid search under the new optimizer.

The best results from the final grid search are listed in Figure 6.



```
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
```

Figure 6: Best Hyperparameters from Grid Search

All models were trained with BCE Loss. For the full hyperparameter results, see
https://github.com/bmanville3/powerstats/tree/main/trained_models

6 Results

Discussion of results with be saved for the next section. We will simply present our findings in this section.

6.1 Training Loss

The graphs in Figure 7 depict the training BCE loss of the best models on the 80/20 training/-validation split from the Open Powerlifting Database.

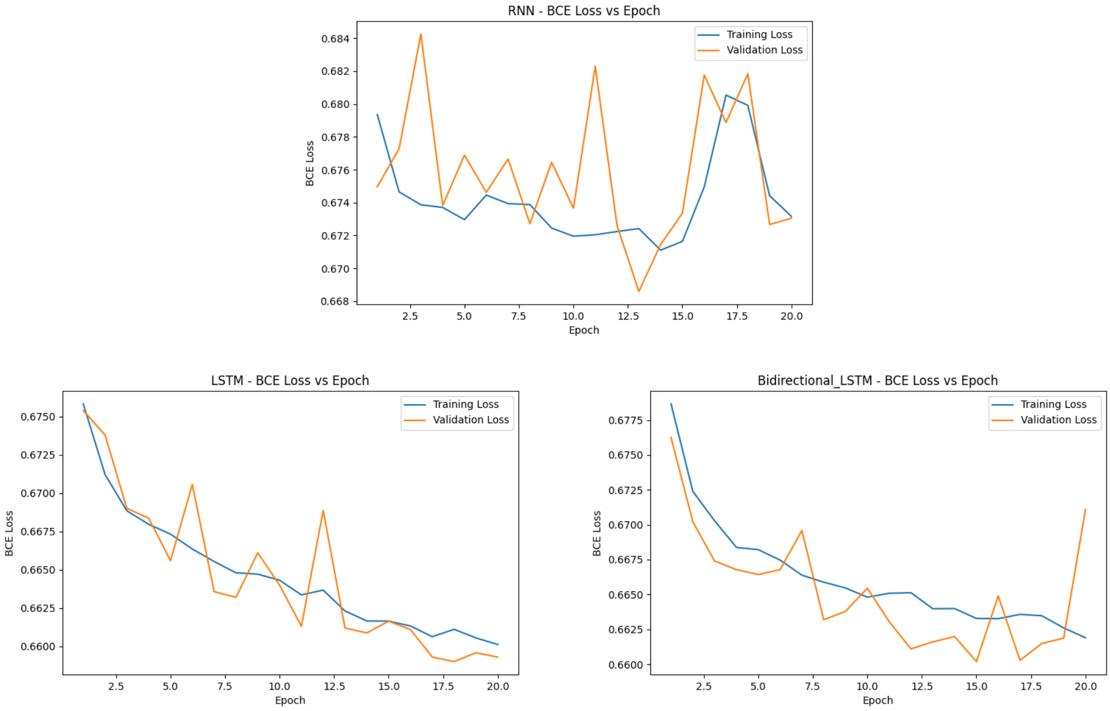


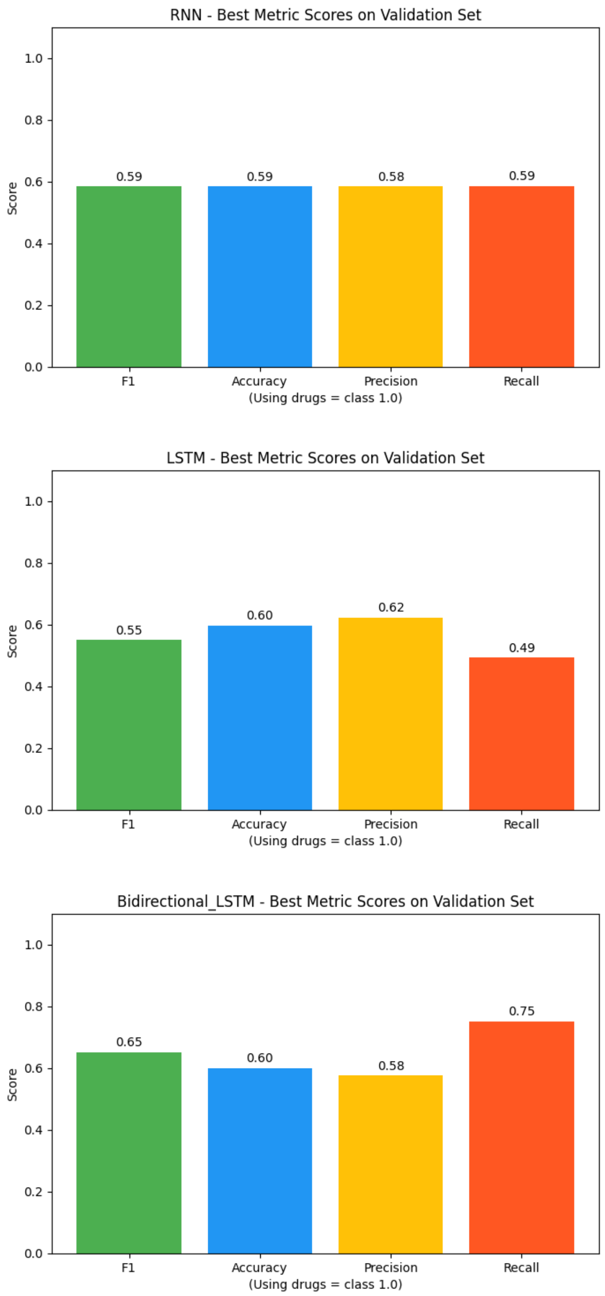
Figure 7: Loss vs Epoch for RNN, LSTM, and BiLSTM

Over the course of 20 training epochs, the binary cross-entropy loss decreased by approximately 0.01 to 0.0175. Given that the initial loss values ranged between 0.67 and 0.68, this represents a relatively small reduction. For reference, a random binary classifier yields an expected loss of around 0.693, so our models did exhibit some learning-but the improvement was limited.

It is worth noting, however, that the initial validation loss is computed after the first epoch of training has already been completed, meaning the true baseline may be slightly higher than reported. Moreover, a high binary cross-entropy loss does not inherently indicate a poor model. Loss should be considered in the context of related work alongside other metrics such as accuracy or F1 score to obtain a more comprehensive evaluation.

6.2 Results on Validation Set

After training the models, the following results were gathered from testing the models on the validation set. Figure 8 displays the results.



RNN - Best Metric Scores on Validation Set

LSTM - Best Metric Scores on Validation Set

Bidirectional_LSTM - Best Metric Scores on Validation Set

Figure 8: Metrics for RNN, LSTM, and BiLSTM on the Validation Set

6.3 Results on USAPL Drug Testing Set

Once all models were trained, we evaluated them on the true test data. Figure 9 displays the results.

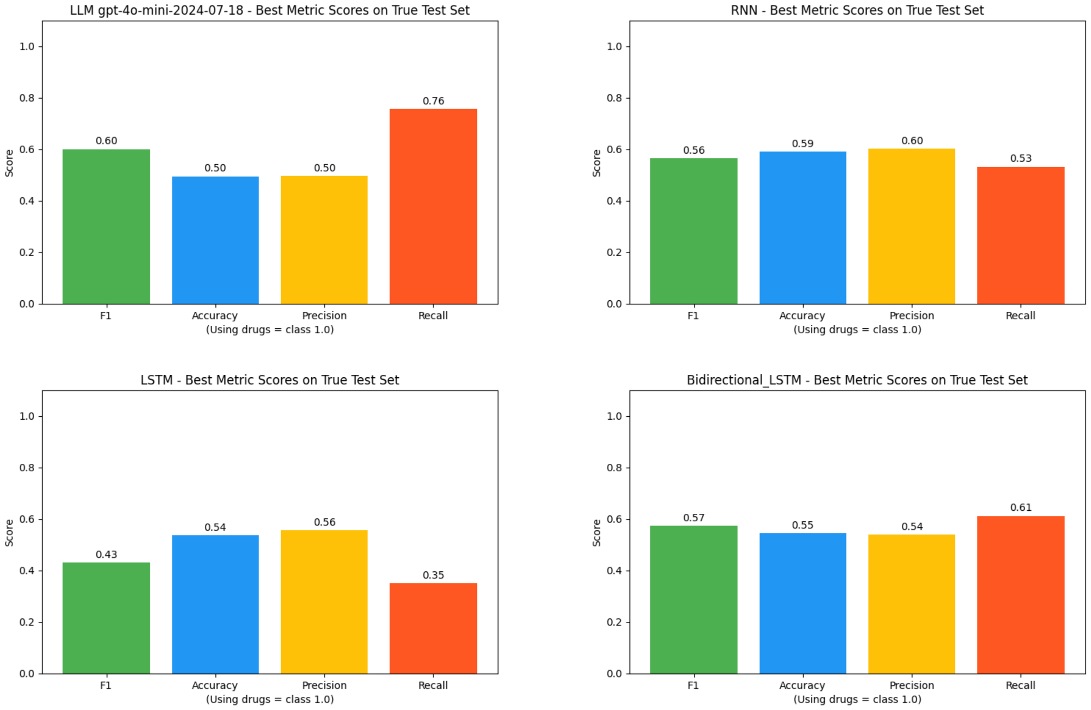


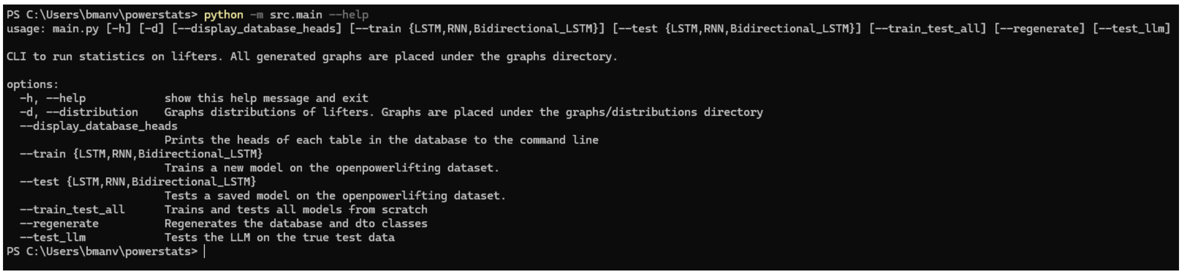
Figure 9: Metrics for LLM, RNN, LSTM, and BiLSTM on the Test Set

To summarize the graphs, the results were as follows:

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

6.4 CLI/GUI

A command line interface (CLI) and graphical user interface (GUI) were made for easy replication, verification, and usage of results. The CLI stands as an interface to train and test new models and generate distributions from the database. The GUI stands as an interface to input data easily, interact directly with the trained models, and query LLMs to generate summaries of lifting data and predict drug use likelihood. Figures 10-11 showcases the CLI and GUI, respectively.

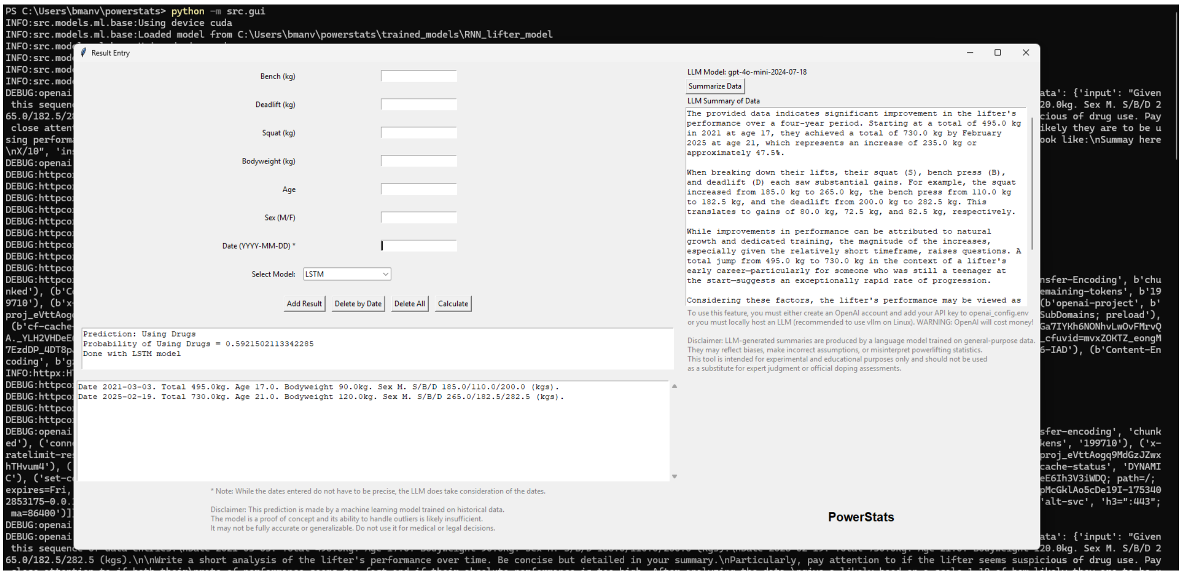


```
PS C:\Users\bmanv\powerstats> python -m src.main --help
usage: main.py [-h] [-d] [--display_database_heads] [--train {LSTM,RNN,Bidirectional_LSTM}] [--test {LSTM,RNN,Bidirectional_LSTM}] [--train_test_all] [--regenerate] [--test_llm]

CLI to run statistics on lifters. All generated graphs are placed under the graphs directory.

options:
  -h, --help            show this help message and exit
  -d, --distribution    Graphs distributions of lifters. Graphs are placed under the graphs/distributions directory
  --display_database_heads
                        Prints the heads of each table in the database to the command line
  --train {LSTM,RNN,Bidirectional_LSTM}
                        Trains a new model on the openpowerlifting dataset.
  --test {LSTM,RNN,Bidirectional_LSTM}
                        Tests a saved model on the openpowerlifting dataset.
  --train_test_all      Trains and tests all models from scratch
  --regenerate          Regenerates the database and dto classes
  --test_llm            Tests the LLM on the true test data
PS C:\Users\bmanv\powerstats> |
```

Figure 10: CLI for Powerstats



```
PS C:\Users\bmanv\powerstats> python -m src.gui
INFO:src.models.ml.base:Using device cuda
INFO:src.models.ml.base:Loaded model from C:\Users\bmanv\powerstats\trained_models\RNN_lifter_model
```

Select Model: LSTM

```
Prediction: Using Drugs
Probability of Using Drugs = 0.5921502113342285
Done with LSTM model
```

```
Date 2021-03-03. Total 495.0kg. Age 17.0. Bodyweight 90.0kg. Sex M. S/B/D 185.0/110.0/200.0 (kgs).
Date 2025-02-19. Total 730.0kg. Age 21.0. Bodyweight 120.0kg. Sex M. S/B/D 265.0/182.5/282.5 (kgs).
```

* Note: While the dates entered do not have to be precise, the LLM does take consideration of the dates.

Disclaimer: This prediction is made by a machine learning model trained on historical data.
The model is a proof of concept and its ability to handle outliers is likely insufficient.
It may not be fully accurate or generalizable. Do not use it for medical or legal decisions.

LLM Model: gpt-4o-mini-2024-07-18

LLM Summary of Data

```
The provided data indicates significant improvement in the lifter's
performance over a four-year period. Starting at a total of 495.0 kg
in 2021 at age 17, they achieved a total of 730.0 kg by February
2025 at age 21, which represents an increase of 235.0 kg or
approximately 47.5%.

When breaking down their lifts, their squat (S), bench press (B),
and deadlift (D) each saw substantial gains. For example, the squat
increased from 185.0 kg to 265.0 kg, the bench press from 110.0 kg
to 182.5 kg, and the deadlift from 200.0 kg to 282.5 kg. This
translates to gains of 80.0 kg, 72.5 kg, and 82.5 kg, respectively.

While improvements in performance can be attributed to natural
growth and dedicated training, the magnitude of the increases,
especially given the relatively short timeframe, raises questions. A
total jump from 495.0 kg to 730.0 kg in the context of a lifter's
early career—particularly for someone who was still a teenager at
the start—suggests an exceptionally rapid rate of progression.

Considering these factors, the lifter's performance may be viewed as
```

To use this feature, you must either create an OpenAI account and add your API key to openai_config.env
or you must locally host an LLM (recommended to use vllm on Linux). WARNING: OpenAI will cost money!

Disclaimer: LLM-generated summaries are produced by a language model trained on general-purpose data.
They may reflect biases, make incorrect assumptions, or misinterpret powerlifting statistics.
This tool is intended for experimental and educational purposes only and should not be used
as a substitute for expert judgment or official doping assessments.

```
65.0/182.5/282.5 (kgs).\n\nWrite a short analysis of the lifter's performance over time. Be concise but detailed in your summary.\nParticularly, pay attention to if the lifter seems suspicious of drug use. Pay
```

Figure 11: GUI for Powerstats

7 Discussion of Results and Conclusions

This project set out to determine whether drug use in powerlifting can be predicted from competition performance data alone. By using a combination of recurrent neural networks (RNNs), Long Short-Term Memory networks (LSTMs), Bi-directional LSTMs (BiLSTMs), and large language models (LLMs), we explored the viability of machine learning for this challenging and underexplored task. Our best model-the classic RNN-achieved an accuracy of 59% and an F1 score of 0.56 on a real-world test set of confirmed doping and non-doping cases, suggesting a modest but non-trivial ability to separate clean and doped lifters using only performance metrics.

To better understand these results, it is helpful to compare them with prior work in adjacent domains. Ryoo et al. (2024), for instance, investigated doping detection in female weightlifters using XGBoost, a Multilayer Perceptron (MLP), and an ensemble of the two [9]. Despite using structured performance profiles derived from athlete passports-arguably more detailed and domain-specific than our raw competition data-they reported only a 53.8% prediction accuracy. Similarly, Hopker et al. (2024) used a Bayesian model to detect suspicious results in elite athletes, offering theoretical support that performance-based signals can reveal doping behavior [8]. However, most of these models require either biological markers, custom feature engineering, or curated athlete profiles, which were not available to us.

In contrast, our approach relied entirely on publicly available competition results and no feature engineering. By framing the prediction task as a proxy classification problem-where the model learns to predict whether a lifter competed in a drug-tested or untested meet-we were able to generate over 300,000 training sequences. While this proxy is imperfect, performance distributions between tested and untested competitions support its use: for example, male lifters in untested meets averaged 569.6 kg total compared to 537.5 kg in tested meets, suggesting a meaningful shift potentially attributable to performance-enhancing drug use.

Among the models trained, the classic RNN emerged as the most reliable, with a 59% accuracy and 0.56 F1 score. Bidirectional LSTMs followed closely, achieving a 55% accuracy and slightly higher F1 score of 0.57. The LSTM alone performed worse (54% accuracy, 0.43 F1), which may indicate that the added complexity of bidirectionality helps capture subtle performance trends over time. Interestingly, the GPT-4o-mini model-evaluated in a zero-shot setting-achieved a 50% accuracy but the highest F1 score of 0.60, indicating that while its predictions were poorly calibrated in terms of accuracy, its balance between precision and recall was strongest. However, this came at a cost: inference took 25 minutes per evaluation and incurred financial expense, whereas RNN-based models completed evaluation in under a minute.

These results collectively suggest that temporal models trained on historical powerlifting results can identify weak but consistent patterns associated with competition drug-testing status. While the gains over random guessing (50%) may seem modest, they demonstrate that doping leaves an imprint on aggregate performance data-an insight that could inform future automated screening tools or probabilistic flags for drug testing committees. In settings where only limited drug testing can be performed, such models could help prioritize athletes for testing based on anomalous performance sequences.

That said, several limitations remain. Most importantly, the test set-sourced from the USAPL drug testing database-contained only 111 confirmed positive and 111 confirmed negative cases. Although balanced, the small size limits generalizability. Moreover, using meet-level drug-testing status as a proxy for individual drug use is inherently noisy, as many lifters in tested meets are not tested, and many untested meets include clean athletes. Despite this, the observed performance differences across divisions support the proxy as a meaningful signal.

In conclusion, while this project does not claim to offer a definitive solution to doping detection, it provides evidence that performance data-when properly structured and modeled-can reveal signals related to drug use. This framework, though preliminary, represents a novel tool in the growing intersection of machine learning and clean sport advocacy.

8 Future Work

Future work should aim to address the limitations that have been outlined in this report. First and foremost, acquiring a larger pool of confirmed drug-testing results would enable more reliable evaluation and potentially support direct supervised learning. Collaborations with federations could make this possible. Second, richer feature representations and data engineering should be done. Incorporating features like DOTs score, Wilks, or other handmand metrics could improve predictive accuracy. Third, integrating biological models or priors could allow for better anomaly detection in uncertain cases. This could also lead to the models being used in conjunction with drug tests to better aid in determining drug use. Lastly, with more data, fine-tuning LLMs could become a viable path, allowing these general-purpose models to better learn domain-specific performance trends.

9 Deliverable and Other Links

9.1 Datasets

Open Powerlifting Dataset:

https://www.openpowerlifting.org/

USAPL Drug Testing Database

https://www.usapowerlifting.com/drug-testing/

9.2 GitHub

GitHub:

https://github.com/bmanville3/powerstats

9.3 Presentation + Demo

Presentation Materials (demo at end in video):

https://docs.google.com/presentation/d/1-ZUHPNLUMNCWcTLasb2LspTu7fJHz6TWzTiPRbeyxsI/edit?usp=sharing

Presentation Video (demo at end):

https://youtu.be/6kNLL7KJAYY

9.4 As Presented in the Introduction

1. A model capable of predicting drug use from performance metrics alone that is better than randomly guessing.

   Deliverable: https://github.com/bmanville3/powerstats/tree/main/trained_models

2. A data analysis and machine learning pipeline capable of training and testing models on powerlifting performance data.

   Deliverable: The entire GitHub repo.

3. A CLI/GUI to access the models and perform anlyses on powerlifting performance data.

   Deliverable (CLI): https://github.com/bmanville3/powerstats/blob/main/src/main.py

   Deliverable (GUI): https://github.com/bmanville3/powerstats/blob/main/src/gui.py

4. An LLM interface to analyze powerlifting performance data.

   Deliverable: Incorporated into the GUI.

10 Skill Learning

10.1 Fact vs Fiction

Throughout the course of this project, I made a consistent effort to separate speculation from evidence. At each checkpoint, I consistently tried to justify my work and reasoning for my actions. Throughout this report, I have done my best to either cite studies for any claims I made or back up the claims with direct evidence I gathered. This skill was particularly important in a domain like drug detection, where direct ground-truth labels are scarce, and it's easy to overstate the reliability of a prediction.

For example, rather than claiming that my model detects drug use directly, I framed the task as predicting whether a lifter competed in a drug-tested meet, and justified this as a proxy. I presented lift total distributions across tested and untested competitions to support the idea that performance-enhancing drugs manifest in observable performance differences. I resisted the temptation to generalize beyond what the data could support and instead highlighted the limitations of the proxy approach.

Additionally, I made my work fully reproducible: code, data cleaning steps, training procedures, and evaluation metrics were all documented clearly, and I published my models and logs to a public GitHub repository. This habit of "showing receipts" has taught me the skill of fact vs fiction.

10.2 Checking Assumptions

This project forced me to routinely examine, challenge, and sometimes discard my own assumptions. Rather than letting assumptions derail progress, I adopted the strategy of "assume temporarily, verify constantly." For example, I made some initial assumptions about data storage at the beginning of the project that started to lead to problems midway through. Rather than pressing on with my original assumption about how the data should be stored (which was too broad), I re-evaluated my architecture and use-cases and came to the conclusion that my original assumptions were false for the scale of my project.

10.3 Trend Recognition

Understanding trends, in both data and model performance, was essential to this project. For instance, I noticed that models consistently hovered around 0.5 confidence for many predictions when I manually tested them in the GUI. This indicated uncertainty and suggested the models weren't overfitting but also weren't fully confident in the separation between tested and untested lifters. At another time, my initial models were all hovering around 60%-70% accuracy before I split the data 50/50 by label. After digging deeper into the metrics the models were producing, I realized the 60%-70% accuracy was coming from the data being 60%-70% tested meets and the models were learning to always predict tested. In another case, by comparing the average lift totals across groups, I observed that male untested lifters outperformed tested ones by a substantial difference that indicated the tested meet proxy may be enough to train models on. Finally, I monitored trends during training, watching how validation loss and F1 scores fluctuated across different configurations. This helped me identify that dropout had limited impact, while hidden layer size and learning rate had much larger effects. Trend recognition was the backbone of my project

10.4 Incremental Progress and Scheduling

In many of my past projects, I had a tendency to procrastinate and would delay most of the work until the final days and rush to complete everything at once. This project helped me break that habit and develop the skill of working incrementally. The regular checkpoint schedule, especially within the condensed timeline of a summer semester along with working full time, required me to make consistent progress week after week. As a result, I learned to break larger tasks into manageable chunks and to schedule my workload more effectively. This incremental workflow not only reduced stress but also improved the quality and consistency of my work.

10.5 Peer Reviewing

Due to the frequent checkpoint submissions, I regularly gave and received peer reviews. This taught me how to to read and evaluate the work of others, track their progress over time, and offer constructive, actionable feedback. It sharpened my ability to compare different approaches and gave me insight into pitfalls and strengths in similar projects. Peer reviewing also served as a reality check for my own progress, helping me gauge how my work compared to my peers' and motivating me to maintain a consistent pace and standard (especially knowing others would be reviewing my work).

10.6 Architecting Scalable and Extensible Code

Although I regularly write scalable and extensible code in my ongoing internship, that work typically involves contributing to a well-established codebase with existing structure and conventions. This project was different in that I had to build the entire architecture from the ground up. It challenged and taught me to think proactively about future needs and modular design. On several occasions, I wanted to add new functionality but found that my earlier code was too rigid or tightly coupled. As a result, I had to refactor parts of the code base. This process significantly improved my ability to design flexible architectures and reinforced the importance of writing maintainable code even in school projects.

11 References

References

[1] Pierre-Edouard Sottas et al. “The Athlete Biological Passport”. In: *Clinical Chemistry* 57.7 (July 2011), pp. 969–976. ISSN: 0009-9147. DOI: 10.1373/clinchem.2011.162271. eprint: https://academic.oup.com/clinchem/article-pdf/57/7/969/32654843/clinchem0969.pdf. URL: https://doi.org/10.1373/clinchem.2011.162271.

[2] *Open Powerlifting*. URL: https://www.openpowerlifting.org/.

[3] James Hopker et al. “Performance profiling as an intelligence-led approach to antidoping in sports”. In: *Drug Testing and Analysis* 12.3 (2020), pp. 402–409. DOI: https://doi.org/10.1002/dta.2748. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/dta.2748. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/dta.2748.

[4] Sergei Iljukov and Yorck O. Schumacher. “Performance Profiling—Perspectives for Anti-doping and beyond”. In: *Frontiers in Physiology* Volume 8 - 2017 (2017). ISSN: 1664-042X. DOI: 10.3389/fphys.2017.01102. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2017.01102.

[5] Sergei Iljukov, Stephane Bermon, and Yorck O. Schumacher. “Application of the Athlete’s Performance Passport for Doping Control: A Case Report”. In: *Frontiers in Physiology* Volume 9 - 2018 (2018). ISSN: 1664-042X. DOI: 10.3389/fphys.2018.00280. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2018.00280.

[6] Raphael Faiss et al. “Editorial: Performance Modeling and Anti-doping”. In: *Frontiers in Physiology* Volume 10 - 2019 (2019). ISSN: 1664-042X. DOI: 10.3389/fphys.2019.00169. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2019.00169.

[7] Silvia Montagna and James Hopker. “A Bayesian Approach for the Use of Athlete Performance Data Within Anti-doping”. In: *Frontiers in Physiology* Volume 9 - 2018 (2018). ISSN: 1664-042X. DOI: 10.3389/fphys.2018.00884. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2018.00884.

[8] 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.

[9] 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.

[10] Hyunji Ryoo et al. "Importance of weightlifting performance analysis in anti-doping". In: *PLOS ONE* 17.2 (2022), e0263398. DOI: 10.1371/journal.pone.0263398.

[11] 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.

[12] C Ayotte et al. "Report to WADA Executive Committe on Lack of Effectiveness of Testing Programs". In: *Montreal: WADA* (2013).

[13] Millán Aguilar-Navarro et al. "Analysis of doping control test results in individual and team sports from 2003 to 2015". In: *Journal of Sport and Health Science* 9.2 (2020), pp. 160–169. ISSN: 2095-2546. DOI: https://doi.org/10.1016/j.jshs.2019.07.005. URL: https://www.sciencedirect.com/science/article/pii/S209525461930095X.

[14] *USAPL Anti-Doping Page*. URL: https://www.usapowerlifting.com/drug-testing/.

[15] Robin M. Schmidt. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. arXiv: 1912.05911 [cs.LG]. URL: https://arxiv.org/abs/1912.05911.

[16] Qingsong Wen et al. *Transformers in Time Series: A Survey*. 2023. arXiv: 2202.07125 [cs.LG]. URL: https://arxiv.org/abs/2202.07125.