Midterm Project Report: Drive Cycle Velocity Prediction

Anwar Rashid Shaik   
 *Data Science*  
*University of New Haven*  
West Haven, CT, USA  
ashai44@unh.newhaven.edu

Vatsalya Nayudu  
Data Science  
*University of New Haven*  
West Haven,CT,USA  
vnayu1@unh.newhaven.edu Vivekananda Reddy Arekatla  
*Data Science*  
*University of New Haven*  
West Haven,CT,USA  
varek1@unh.newhaven.edu

Manogna Vennela Ramireddy  
Data Science  
*University of New Haven*  
West Haven,CT,USA  
mrani7@unh.newhaven.eduPavan Reddy Boyapally  
Data Science  
*University of New Haven*  
West Haven,CT,USA  
pboya2@unh.newhaven.edu

*Abstract*—The project on Drive Cycle Velocity Prediction aims to develop a model capable of predicting vehicle speed patterns based on past data. This work combines time-series analysis and machine-learning techniques using standard and synthetic datasets to produce accurate, time sensitive velocity predictions. Furthermore, the project evaluates conditional probabilities of reaching different speed levels within specified windows, based on the current speed. These findings have potential applications in automotive testing, fuel efficiency analysis, and emissions studies.

Keywords—Drive Cycle Prediction, Time-Series Analysis, Velocity Prediction, Conditional Probability, LSTM, Machine Learning, Synthetic Data

# Introduction

In automotive engineering, accurate velocity prediction is crucial for assessing fuel economy, emissions, and overall performance under different driving conditions. Drive cycles, which are standard tests simulating real-world driving behaviors, provide the necessary data to model these dynamics. This project utilizes both standard and synthetic drive cycle datasets to develop a time-series model that can accurately predict vehicle velocities.

***Literature Review***—Existing studies have applied machine learning techniques, particularly recurrent neural networks such as LSTMs, to predict speed patterns based on time-series data. The availability of drive cycle datasets from organizations like the US Environmental Protection Agency (EPA) and various European standards provides a solid foundation. However, a lack of variability within standard datasets has prompted the addition of synthetic datasets with controlled noise. Our approach combines both data types to build a model that not only predicts speed accurately but also evaluates conditional probabilities of speed transitions over specified windows.

# Objective

The primary objective of this project is to develop a model capable of predicting vehicle velocities based on past driving data. The specific goals are as follows:

* Train a time-series model on both standard and synthetic datasets to predict velocity.
* Compare the predicted velocities against actual values using histogram analysis.
* Evaluate conditional probabilities to predict speed changes within a window, given the current speed.

# Methodology

## Dataset Collection and Preprocessing

The project began with the collection of drive cycle datasets, including both standard and synthetic data. The standard datasets were sourced from recognized repositories such as the US Environmental Protection Agency (EPA) and other European driving cycle datasets. These datasets represent realistic driving conditions, including speed variations that occur during acceleration, cruising, deceleration, and idling phases.

To enhance variability and account for different driving behaviors not captured in the standard data, synthetic datasets were generated. Synthetic data was created by adding controlled noise to the standard datasets. This noise was added within a specific range (e.g., ± 3 mph) to maintain the integrity of realistic speed transitions while introducing variations to increase the robustness of the prediction model.

The collected datasets were divided into training and validation sets. Each dataset underwent preprocessing to remove any inconsistencies, such as missing values or outliers, that could negatively impact model performance. The preprocessing steps also involved normalizing the speed data to ensure that all values were on a consistent scale, which is essential for effective model training. Additionally, time-series data were segmented into smaller sequences to serve as input-output pairs for the model.

## Model Development

The project employed two key approaches for developing predictive models: the Velocity Prediction Model using LSTM and the

Conditional Probability Analysis for understanding speed transitions.

### Velocity Prediction

* Model Selection: The Long Short-Term Memory (LSTM) network was chosen for its ability to effectively capture temporal dependencies in sequential data. LSTMs are well-suited for time-series analysis, which is essential for predicting future vehicle speeds based on past data. Unlike traditional feedforward neural networks, LSTMs contain memory cells that allow them to retain information across long sequences, making them ideal for modeling the complexities of vehicle acceleration and deceleration.
* Architecture: The LSTM model architecture included one LSTM layer with 50 units, followed by a dense layer that outputs the predicted speed. The number of units was selected after a series of hyperparameter tuning experiments to determine the optimal model configuration. The Adam optimizer was used for training the model, along with the Mean Squared Error (MSE) loss function to minimize prediction errors.
* Training Process: The model was trained on both standard and synthetic datasets to improve generalizability. During training, the model learned from sequences of past vehicle speeds to predict the next speed value in the sequence. Training involved multiple epochs to ensure convergence, and the batch size was carefully selected to balance training efficiency and model accuracy. Cross-validation was also performed to ensure that the model was not overfitting to any particular dataset.
* Evaluation Metrics: The model's performance was evaluated using Mean Squared Error (MSE), which measures the average squared difference between predicted and actual speed values. Additionally, histograms comparing predicted and actual speeds were used to validate that the model captured the overall distribution of speeds accurately.

## Data Augmentation and Synthetic Data Generation

Given that real-world driving data can often be limited or lack variability, data augmentation played a key role in this project. By creating synthetic datasets through noise addition, we introduced variability that might be encountered in different driving environments or under varying driver behaviors. Each team member was responsible for generating two synthetic datasets from two standard datasets, resulting in a total of ten synthetic datasets. The process involved adding controlled random noise while ensuring that the resulting speed values remained realistic and that the driving patterns (e.g., acceleration and deceleration phases) were preserved.

The synthetic datasets were validated by comparing their histograms and time-series plots against those of the standard datasets to ensure that the introduced noise did not significantly distort the overall driving patterns. This step was crucial for maintaining the integrity of the training data and ensuring that the model could learn effectively from both real and augmented data.

## **Training and Model Optimization**

The training process involved iterative model development, where multiple versions of the LSTM model were trained, evaluated, and refined. Hyperparameter tuning was conducted to identify the best combination of parameters, such as the number of LSTM units, learning rate, and batch size, that would yield the lowest validation error. To prevent overfitting, regularization techniques such as dropout were applied, and early stopping was used to halt training when the validation performance ceased to improve.

The model was trained on GPU-accelerated hardware to expedite the process, given the computational intensity of LSTM networks. Each training iteration involved monitoring the loss curves and adjusting the model parameters to ensure that both training and validation losses were minimized without significant divergence, indicating good generalization to unseen data.

## Individual Contributions

1. *PAVAN:*

* Role: Data preprocessing and literature review, synthetic data generation.
* Contributions: Conducted literature review on existing methods, collected and prepared datasets for model training, and generated 2 synthetic datasets from 2 standard datasets.
* Results: Preprocessed both standard and synthetic datasets, making them suitable for LSTM input. Created 2 synthetic datasets for use in model training.
* Future Tasks: Fine-tune model parameters and further refine data for the final model.

1. *ANWAR RASHID SHAIK:*

* Role: Model training and evaluation, synthetic data generation.
* Contributions: Trained the velocity prediction model on standard and synthetic datasets, performed initial evaluations using MSE, and generated 2 synthetic datasets from 2 standard datasets.
* Results: Achieved baseline MSE and created initial model graphs for time-series prediction. Created 2 synthetic datasets to support model training.
* Future Tasks: Improve the model’s accuracy and compare histogram results for further analysis.

1. *VIVEK:*

* Role: Synthetic data generation and conditional probability analysis.
* Contributions: Generated 2 synthetic datasets from 2 standard datasets and analyzed conditional probabilities for speed changes.
* Results: Developed preliminary probability models, created 2 synthetic datasets, and explored different probability thresholds.
* Future Tasks: Complete the probability model and finalize synthetic data for improved model training.

1. *VENNELA:*

* Role: Model evaluation and documentation, synthetic data generation.
* Contributions: Evaluated model performance using various metrics such as MSE and accuracy, documented results, assisted in creating visualization tools for analysis, and generated 2 synthetic datasets from 2 standard datasets.
* Results: Compiled initial performance metrics and visualizations for time-series and histogram comparisons. Created 2 synthetic datasets.
* Future Tasks: Refine model evaluation techniques and prepare detailed documentation for final report submission.

1. *VATSALYA:*

* Role: Data visualization and report compilation, synthetic data generation.
* Contributions: Developed visualizations for the velocity prediction model, including histograms and time-series plots, assisted in compiling the midterm report, and generated 2 synthetic datasets from 2 standard datasets.
* Results: Created effective visual tools to compare actual vs. predicted values, contributed to clearer communication of project outcomes, and created 2 synthetic datasets.
* Future Tasks: Continue enhancing visualizations and contribute to the final project report preparation.

# RESULTS

## VELOCITY PREDICTION MODEL

### Mean Squared Error (MSE): The model achieved an initial MSE of [X value]. This demonstrates the model’s current performance in predicting the velocity sequence based on past values.

### Histograms: Histograms comparing actual and predicted velocities show that the model successfully captures the general distribution and trends of the data.

### *Time-Series Comparison:* Time-series plots illustrate how well the model’s predictions align with the actual velocity trends over time, with periodic adjustments to fine-tune accuracy

# Conclusion

This project has successfully developed a foundational velocity prediction model using both standard and synthetic data, leveraging time-series analysis for accurate forecasting. Preliminary results demonstrate the effectiveness of the LSTM model.

*Future Work:*

Further refinements will focus on optimizing model accuracy, exploring additional time-series architectures to enhance the model’s robustness across different driving scenarios. Additionally, we plan to:

* *Integrate the velocity prediction model with electric vehicle controllers:* Using the MATLAB simulation, we will modify and optimize the existing control strategies based on our velocity predictions. This will involve adjusting power management and control parameters for improved energy efficiency and smoother driving experience.
* *Use synthetic datasets for controller testing:* The synthetic datasets will be used to test the robustness of electric vehicle controllers under various simulated driving scenarios, ensuring adaptability to different conditions.
* *Explore the integration of Model Predictive Control (MPC):* We will explore integrating our LSTM-based velocity prediction model within an MPC framework to enable the controller to make predictive decisions based on future states, improving vehicle dynamics and efficiency.