In softball, the strike zone isn't as straight forward as it appears to be in the rulebook. The NCAA defines the strike zone as the area between the bottom of the batter's sternum and the top of her kneecaps, while in a natural stance. The strike zone also spans the entire 17-inch width of home plate, meaning that any part of the ball that crosses home plate is eligible to be called a strike. This leaves a lot of room for individualized interpretation from umpires based on a variety of factors such as pitch location, count, pitcher and batter handedness, and game context. To study this, we built a called strike probability model that predicts how likely a taken pitch, meaning there was no swing, is to be called a strike. We originally trained the model on MLB Statcast data from the 2020-2024 seasons and later adapted it to LSU TrackMan data from the 2025 season. The model uses machine learning to look at pitch type, count, and handedness, so we can understand how each of these factors affects the zone and influences pitch and swing decisions. And overall, these results can help improve training, player evaluation, scouting, and in-game decision-making for LSU Softball.

We started with the MLB Statcast data because the LSU TrackMan data wasn't available at the beginning of the semester. The Statcast dataset included everything we needed: pitch location at the plate, pitch type, velocity, count, pitcher handedness, hitter handedness, pitch result, and more. For our purposes, we filtered this dataset down to taken pitches only and defined our response variable as 1 for a called strike and zero for everything else. Later in the semester, we received the LSU TrackMan data, but this dataset was much smaller than the MLB dataset. It contained the key variables we needed like PlateLocSide, PlateLocHeight, pitch type, count, pitcher and batter handedness, and the swing indicator. We cleaned this dataset by removing any rows with missing pitch-location values and keeping only the variables needed for the model.

We started by studying the MLB pitch locations so we could understand what a typical strike zone pattern looks like in a large, stable dataset. We then generated heat maps to show the distribution of taken pitches and where umpires most often call strikes. Upon receiving the LSU Softball data, we generated new heat maps to show the raw pitch tendencies before any modeling. For example, with curveballs, left-handed pitchers show wider horizontal variability, while both RHP and LHP curves tend to finish low in the zone. These shapes look different from the MLB patterns because softball has different pitch types with different release points. Additionally, the LSU dataset is much smaller, which affects the density and smoothness of the heat maps.

In terms of model architecture and methods, we first filtered the new softball dataset to remove any irrelevant or incomplete data. This gave us a clean, consistent foundation to work with. Next, we adapted our original MLB-trained model to fit the new softball data. To do this, we expanded the model’s input layer from two neurons to four, allowing us to include two new variables: pitcher handedness and batter handedness. These features are important in softball because they change pitch movement patterns, affecting the way umpires call the zone. Instead of retraining a new model from scratch, we built on our existing MLB-trained model using neural network weight surgery. We copied the original first layer weights into a large-weight matrix then set the new columns to zero. This keeps the model’s original behavior completely intact at the start because the new inputs do not influence any predictions yet. From there, we fine-tuned the model on the softball data, letting it gradually learn how handedness affects pitch outcomes while still receiving help from everything it learned from the MLB data. Although the current dataset is limited, the trends are clear, and the model provides a practical foundation for training applications. Future work will involve collecting more LSU softball seasons, incorporating additional features (framing, swing behavior, and umpire tendencies), and building individualized strike zone visualizations for players and umpires. These additions will improve accuracy and strengthen the model’s usefulness for preparation, development, and in-game decision support.