# Investigating Explainability of ML Model Decisions for Injury Surveillance

Mentor: Romila Pradhan rpradhan@purdue.edu
Mentor: Gaurav Nanda gnanda@purdue.edu
Manas Paranjape mparanjape@purdue.edu
Harmya Bhatt hvbhatt@purdue.edu

## Introduction & Motivation

- One of the applications of machine learning (ML) in public health and safety is to assign diagnostic codes to injury narratives recorded at hospital emergency rooms or reported by employers and insurance companies
- It is often difficult to understand why a machine learning model predicts a particular code for an incident
- We aim to identify parts of the injury narratives most responsible for the ML model outcomes using advances made in Explainable Artificial Intelligence (XAI)
- We will generate explanations for overall model behavior and model decisions for particular narratives
- This project is part of our efforts to generate human-understandable explanation model outcomes deployed in the areas of public health and safety, and will be instrumental in advancing the state of the art in the automated encoding of safety data

## Methodology

- We have used a dataset by Occupational Safety and Hazard Administration (OSHA) to train our model
- We are predicting on 2 classes, falls to lower level and falls on same level
- We are splitting the data into training and testing data with an 80%-20% split
- We are using a Logistic Regression classifier with accuracy as a performance metric

### LIME

LIME [1], or Local Interpretable Model-Agnostic Explanations, is an algorithm that can explain the predictions of any classifier or regressor.

Original Image

Interpretable Components

- LIME produces an explanation by approximating the original model by an interpretable model in the neighborhood of the instance we want to explain
- We use the output of LIME to explain cases where the model correctly predicts an event, and more importantly, where the model incorrectly predicts an event

The explainable model is usually exploited to understand which variables are the most important for the ML prediction for the specific event narrative

### SHAP

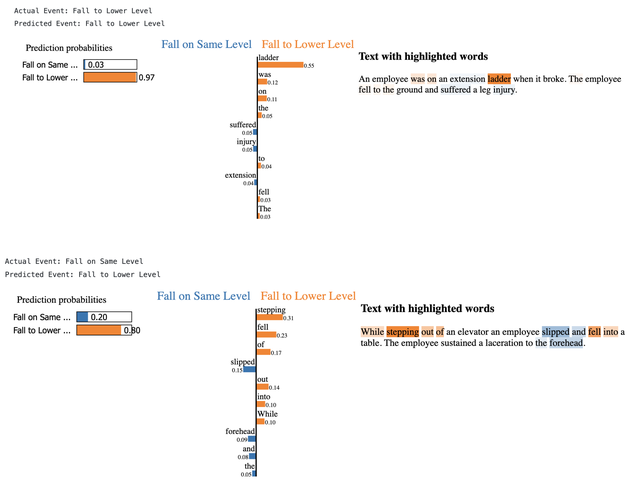
SHAP [2], or SHapley Additive exPlainations is a game theory approach to explain machine learning models. It helps determine the reasoning behind a machine learning model predictions. SHAP explains models using the following methodology:

- Shapley values are calculated for various features in the model using cooperative game theory
- These Shapley values tell us how to fairly distribute the "payout" (in this case, prediction) among the features
- In our case, individual words or sequences of words are considered as features
- We can plot these Shapley values in order to get a better visual understanding of the model and its prediction on certain data

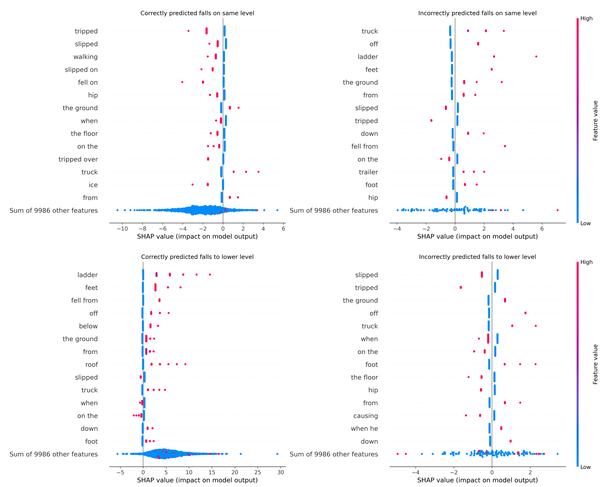
## Results & Discussions

- The model we built had an accuracy of 94% on average in a 5-fold cross validation
- However, it still confused outcomes between *fall on same level* and *fall to lower level*
- to explain this, we decided to use SHAP and LIME to explain the reasons behind these errors

### LIME




### SHAP




### LIME

- The above graphs indicate a local interpretation of two cases: a case where the model **correctly** predicts the event and a case where the model **incorrectly** predicts the event
- In the first case where *the model* ***correctly*** *predicts the event*, features like *'ladder'* and *'on'* pushed the model towards *falls to lower level*
- In the case of *the model* ***incorrectly*** *predicts the event*, features like *'stepping'* and *'fell'* confused the model by pushing it towards *falls to lower level* and away from *falls on same level*

### SHAP

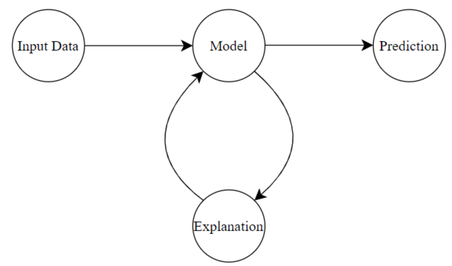
- The above graphs provide a comparative study of which features contribute to the **incorrect** and **correct** predictions for each class
- In the case of *falls on same level*, features like *'ladder'* confused the model and pushed the model towards *falls to lower level*
- In the case of *falls to lower level*, features like *'tripped'* increased the chances of incorrectly predicting *falls on same level*

## Conclusions

- The results demonstrate that there are various features which confound the model causing incorrect predictions
- These confounding features are often important features to successfully predict the other class as they push the result in the opposite direction
- Use of these features in specific context may cause these errors as Logistic Regression model does not take into account relative positioning of words
- Inducing context into the data or model may help improve model accuracy

## Future Work

- We plan to explain the models specifically for incorrect predictions and find the most important factors driving these faults
- Based on this, we aim to find the features which are repeatedly pushing the predictions towards false negatives or false positives
- In order to reduce this confusion, we plan to use n-grams to determine which sequences of words are important in predicting the outcomes




[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 3016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).
[2] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)