Stat 380 Final

Patrick Erickson, Aiden Shindel, Jordan Brophy

Predicting Drowsy Driving: Can we Machine Learning Model to be able to Predict Drowsiness in Drivers?

Abstract

In this work, the main objective is to develop a system that can detect fatigue of any human and issue a timely warning. Drivers who do not take orderly breaks while driving for long distances run into a high chance of turning drowsy, which deteriorates the ability to drive safely. We address the challenge of detecting driver drowsiness from facial images by developing a highly accurate prediction system. We address the challenge of detecting driver drowsiness from facial images by developing a highly accurate prediction system. We first applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance the images, improving both human and computational interpretation. This technique removed noise and defined repetitive patterns, ensuring that no features were overlooked due to low contrast, exposure, or other image quality issues. In addition, we used High Frequency Emphasis to exacerbate the defined edges even further. Principal Component Analysis (PCA) was then applied to identify key patterns associated with yawning and drowsiness while reducing data dimensionality for more efficient modeling. We trained 12 distinct models with an ablation of these pre-processing and feature extraction techniques across three architectures: Gaussian Support Vector Machines (SVM), Random Forests (RF), and XGBoost using 5 fold cross validation combined with a rudimentary hyperparameter search. To further enhance predictive performance, we developed a Logistic Regression Stacking Classifier, strategically integrating the strengths of each model: margin maximization from Linear SVM, non-linear pattern detection from XGBoost, and robustness from Random Forests. Our final ensemble achieved an overall accuracy of 96.9% via logistic regression, significantly outperforming the best individual model which achieved 86.2% accuracy.

Introduction

Drowsiness is a process in which the level of consciousness is reduced due to lack of sleep or fatigue and it may cause the driver to fall asleep. When the driver is suffering from drowsi-

ness they lose control of the car, so they might suddenly deviate from the road and hit an obstacle. Drowsy driving has become a systemic issue in the United States rooted in the busy nature of daily life that is leading to more deaths every year. Unfortunately, determining a precise number of drowsy-driving crashes, injuries, and fatalities is not yet possible. Crash investigators can look for clues that drowsiness contributed to a crash, but these clues are not always identifiable or conclusive. NHTSA estimates that in 2017, 91,000 police reported crashes involved drowsy drivers in the United States. These crashes led to an estimated 50,000 people injured and nearly 800 deaths. Although there is awareness for driving under the influence of specific medications and impairments, drowsy driving is not talked about nearly enough. Therefore, we will use machine learning tools to accurately sort images of human faces into drowsy and natural categories. Unlike traditional statistical models, which often rely on predefined assumptions, machine learning methods are designed to adapt and improve as they're exposed to more data. This flexibility has made them particularly useful in settings where data is messy, highly dimensional, or constantly changing. Because modeling a dataset of images can be more difficult than quantitative variables, machine learning must be used in order to create the best possible model.

Literary Review

The AAA Foundation for Traffic Safety investigated the true impact of drowsy driving on vehicle crashes in the United States from 2009 to 2013.[6] While official government statistics estimate that drowsiness contributes to only 1–3% of crashes, this study suggests the real numbers are much higher. Using detailed crash investigations and statistical imputation techniques, the study found that approximately 6% of all crashes involving towed vehicles, 7% of injury crashes, 13% of hospitalization crashes, and 21% of fatal crashes involved a drowsy driver. Nationally, this equates to about 328,000 crashes each year, resulting in roughly 109,000 injuries and 6,400 deaths annually. The study emphasized that traditional police reports often fail to capture drowsiness because it leaves no obvious physical evidence, unlike alcohol impairment. Many drivers either do not realize they were drowsy or are unwilling or unable to report it. By using a representative sample from the National Automotive Sampling System (NASS) and applying multiple imputation methods, the researchers were able to estimate the prevalence of drowsiness even when the driver's alertness was officially recorded as "unknown." Overall, the findings highlight that drowsy driving is a far more serious and under-recognized threat to traffic safety than previously believed, especially in severe and fatal crashes. The study titled "Heterogeneous ensemble learning for enhanced crash forecasts" explores the application of ensemble machine learning techniques to improve the prediction accuracy of crash frequencies on roadway segments.[1] By leveraging multiple diverse models, the research aims to provide more reliable forecasts of future roadway safety, which can inform better infrastructure planning and traffic safety interventions. The findings suggest that heterogeneous ensemble learning approaches can significantly enhance the precision of crash predictions compared to traditional single-model methods.

Methodology

The dataset used for this project is available on Kaggle courtesy of Jebrailey et. al.[4]. The original kaggle URL can be found here: https://www.kaggle.com/datasets/yasharjebraeily/drowsy-detection-dataset. This dataset adheres to the FAIR CARE principles by being Findable, Accessible, Interoperable, and Reusable. It provides labeled images of human faces under different states of drowsiness, making it easy to access and work with. The dataset is well-structured, with proper metadata, which ensures it can be reused effectively for future research and applications in the domain of drowsiness detection.

Ethical Considerations in the Choice of our Data Set

In order to ensure the ethical feasibility of the data set, we also scrutinized our selection to follow the F.A.I.R and C.A.R.E principles. Since we had obtained our data from the open-source data hub Kaggle, we can ensure that anyone can find and access this data set through Kaggle's use guidelines, stating for a free distribution of data sets that are published for public use to practice ML or data analysis on. The data is also highly versatile, taking a csv format. This is one of the most interoperable formats there are for data sets, especially for our use-case in R, where reading in a csv is already built into the base. Lastly, the data set can be reused for multiple types of data analysis, and will always remain relevant for its time period. In terms of collective benefit, the data set contributes to an overall understanding of of drowsy dribing without harming any individuals or teams, due to the fact that it is simply raw data. Secondly, as I had stated the author's name, there is clearly ownership over the data set, showing proper Authority to Control. Since proper attributions are maintained, it also falls under the Responsibility and Ethics guidelines of the C.A.R.E principles.

Our methodology pipeline can be outlined as such:

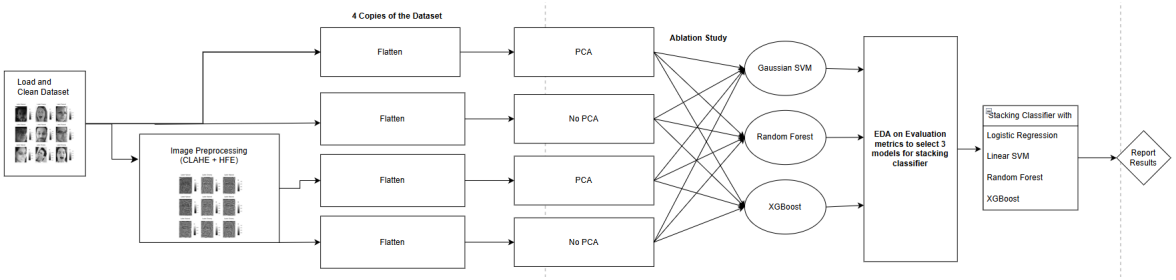

Figure 1: This is a diagram of our methodology. The following will delve more into detail about each constituent part.

Cleaning

To clean the Drowsy Drivers dataset, we first filtered out all .jpg file types, retaining only .jpg_0.png and .png files. The dataset was then combined into a single training and test set. To ensure consistency, we shuffled the dataset and ensured that both the training and test sets were balanced. We performed an ablation study to assess the impact of various features. The dataset was split into 80% for training, 10% for validation and training the meta models, and 10% for testing to ensure proper model evaluation, avoiding data leakage. We then make copies of the dataset, using CLAHE and HFE to apply pre-processing.

CLAHE

The first augmentation technique we use therefore is Contrast Limited Adaptive Histogram Equalization (CLAHE). This is a refined version of normal Histogram Equalization that contains a regularization parameter to mitigate the strength of the initial histogram equalization technique by adding a clipping feature to the initial dataset, ensuring that contrasts are never set too high. Furthermore, the image is subdivided into separate sections, where histogram equalization is applied separately. These sections are then combined using bilinear interpolation, which ensures a smooth transition between each segmented histogram. This way, we do not create entire areas of extreme contrast and intensity, something that normal histogram equalization is prone to doing. We therefore opted to maintain this augmentation technique in our pipeline. Lastly,

HFE

High-Frequency Emphasis (HFE) is a type of processing technique that boosts subtle details and fine textured by enhancing the frequency components of a given image. The "frequency" is usually represented via a Fourier Transform into many different frequency elements, where low frequencies correlated to slowly varying information (think features that are fairly consistent throughout an image) and high frequencies relating to contrasting sections of an image (areas where features abruptly change). This allows for the convolutions in computer vision to better pick up patterns and improve generalizability. After carefully analyzing our own parameters and the differences between our and our parents' methodologies, we opted to maintain this data processing technique, as it is vital for the extrapolation of the plant diseases. For our high frequency emphasis, we use the following Gaussian blur for convolutions with a sliding of 1 pixel per convolution.

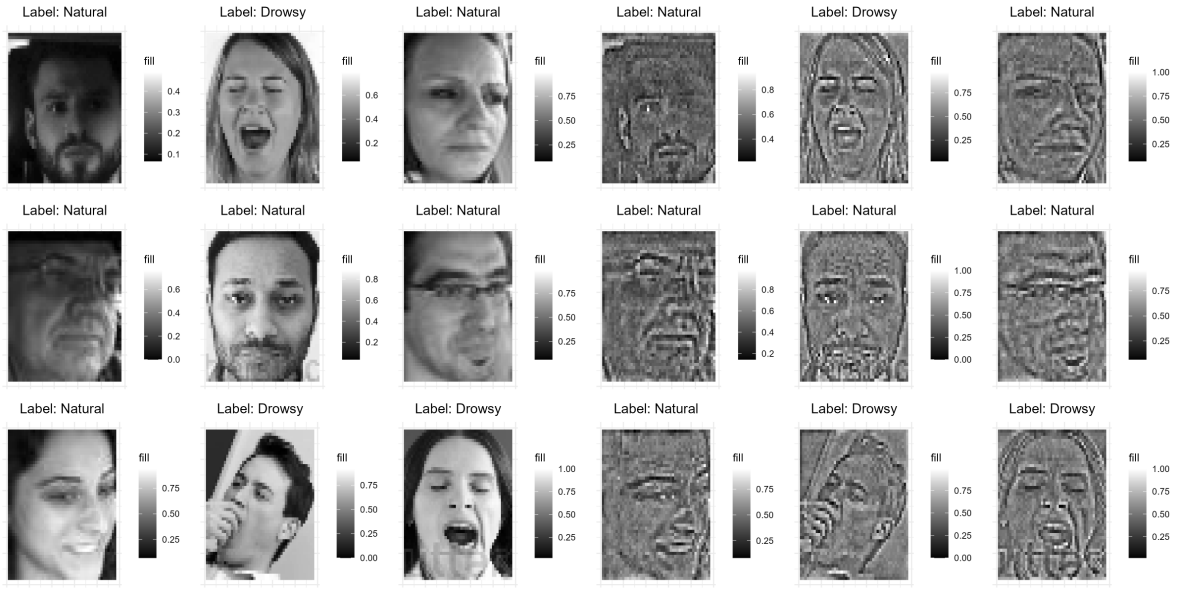

Figure 2: On the left we have our original images. On the right is the CLAHE + HFE transformed images. Notice how the contrast of the edges are emphasized and the noisier features become more smooth in represenation.

Feature Extraction: PCA

To optimize drowsiness detection, we applied Principal Component Analysis (PCA) to reduce the high dimensionality of our image data while preserving the most important recurring patterns. By capturing 95% of the total variance, PCA allowed us to retain the essential structure of the data with far fewer components PCA was used to reduce the number of input features, helping simplify the data, remove noise, and speed up training especially important for models like Support Vector Machines (SVMs). We chose to compare these additions of pre-processing and feature extraction techniques to ultimately see if we could gain valuable information from some subset of feature engineering, no feature engineering, PCA, and no PCA. By skipping PCA and pre-processing for them, we allowed our models to access the full richness of the original enhanced image data, potentially capturing subtle patterns that PCA might compress away, or that CLAHE and HFE may obscure.

Flattening

Next, our dataset was flattened. This was done so the data is transformed from the multidimensional feature space into a one dimensional vector. This process involved taking the original 48x48 pixel matrices and concatenating them into a single feature vector for each image consisting of the variables corresponding to the particular subset of pre-processing and

feature extraction techniques on the dataset. Flattening the data allowed us to feed the output into machine learning models, which required a one-dimensional input, ensuring compatibility and improving the efficiency of the model training process.

Model Selection and Training

In the context of drowsiness detection, we decided to stick with classifiers that were able to capture feature information in high-dimensional spaces and be able to construct non-linear decision boundaries due to the difficulty in classifying images by their pixel information. Therefore, random forest was suitable because it handles many input features as after PCA and captures complex, nonlinear patterns in the facial data. Each tree in the forest can split on different aspects of the face such as eye closure metrics or grayscale values to decide if the driver is drowsy or alert. The ensemble nature of random forests also makes them robust to noise and missing data; if some pixel features are unreliable due to shadows or occlusions, other trees using different features can compensate. Moreover, random forests naturally output feature importances, which could help us interpret which facial features or principal components are most indicative of drowsiness. XGBoost in particular is known for its efficiency and accuracy. It incorporates regularization or penalties on tree complexity and advanced optimizations such as parallel tree construction and cache aware data structures to prevent overfitting and handle many features. For drowsiness detection, XGBoost can be very effective because it continuously refines the model's focus on hard examples such as drivers whose eye features are borderline between alert and drowsy. By adjusting subsequent trees to reduce those misclassifications, XGBoost often achieves higher accuracy on complex tasks. In practice, our XGBoost classifier would take the PCA reduced features and build a series of decision trees that, together, form a highly tuned predictor of the driver's state. Gaussian SVMs are particularly effective in high dimensional spaces and are robust to overfitting when only a subset of points matter. This suits our problem because the PCA step can produce a moderate number of features even from high resolution images, and we may have a limited number of training samples. In essence, the SVM treats the features of each face as a point in gaussian kernel feature space and learns a boundary between "drowsy" and 'natural" clusters, maximizing separation. Because SVM focuses on the hardest to classify points and handles complex, potentially overlapping classes, it can capture subtle distinctions in facial geometry or texture that indicate drowsiness. In our experiments, we used the radial basis function kernel in caret in R to accomplish this task. The downside is that SVM training can be slower than tree methods for large datasets, but after PCA the feature space was compact enough to make SVM feasible, and tuning its regularization helped avoid overfitting. We performed an ablation study with all of these models in order to construct an empirical analysis for our stacking classifier.

Stacking Selection and Training

After training all three models, we combined their strengths via a stacked ensemble using simple models such as logistic regression and linear SVM, as well as more complex models such as Random Forest and XGB as the meta learner. In stacking, the predictions of the base models become inputs to a higher-level model that learns how to best combine them. Here, the best three base models from the previous section were chosen based on an empirical analysis, where we determine which of the base models have the best performance while covering unique strengths. Each of these three models produces a predicted class for an image, and the meta learner learned the optimal weights to weight these predictions. Conceptually, stacking harnesses the "wisdom of crowds": by using multiple diverse classifiers, the ensemble tends to outperform any single one. Empirically, our logistic regression stacking indeed improved overall performance on the drowsiness task. As noted in ensemble learning literature, combining several models can reduce both bias and variance: if one model underfits and another overfits, the stack can balance them, and inconsistent errors tend to cancel out. In our case, random forest provided robust, variance-reduced predictions, XGBoost contributed finely tuned decision rules for difficult cases, and SVM added a large-margin perspective. By training a new model on their outputs, the ensemble effectively interpolated among these strategies.For ALL models, both the base models and the stacked classifier, 5-fold cross-validation with a basic hyperparameter search was performed.

A Caveat: Interpretability

Due to the fact that feature extraction for 2 dimensional images may have the most prominent features spread about different features due to some affine transformation in the 2D space provided, individual feature interpretability (representing a pixel) is extremely diminished, and very little information can be extracted from any given feature. Furthermore, due to the fact that we are maximizing accuracy, our interpretability will come from the empirical analysis of the ablation study, where we will be pitting different models against each other and suggesting what each model may have "learned" based on the training of the model. We will use this to also explain the effect of "stacking". In order to maintain some semblance of interpretability, we will be maintaining a small number of models in our stacking classifier to aid in visualization. Below is an example of what this "Wisdom of the Crowds" phenomena would look like.

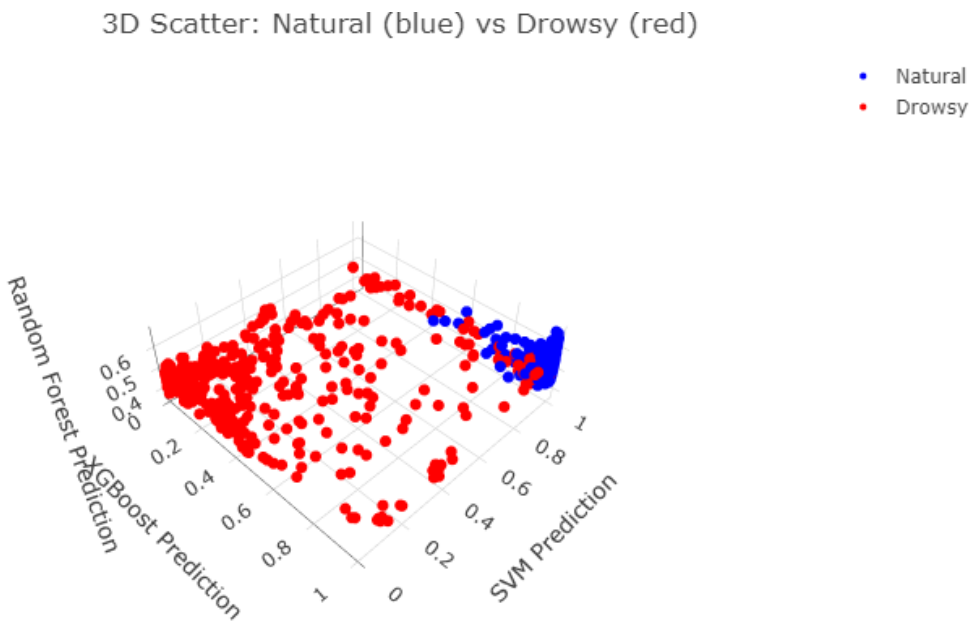

Figure 3: You can see that with the predictions from the models, we can build a new feature space where the labels are able to be separated much easier.

Results

Without any pre-processing, the images were reduced to 183 principal components. After applying pre-processing techniques such as automatic CLAHE and High Frequency Emphasis, a much larger 1,313 components were required to capture the same level of variance, reflecting the increased complexity and richness of the enhanced images. Although using a higher number of components made the models more computationally intensive, it was a necessary trade-off to ensure that critical patterns related to drowsiness were accurately recognized.

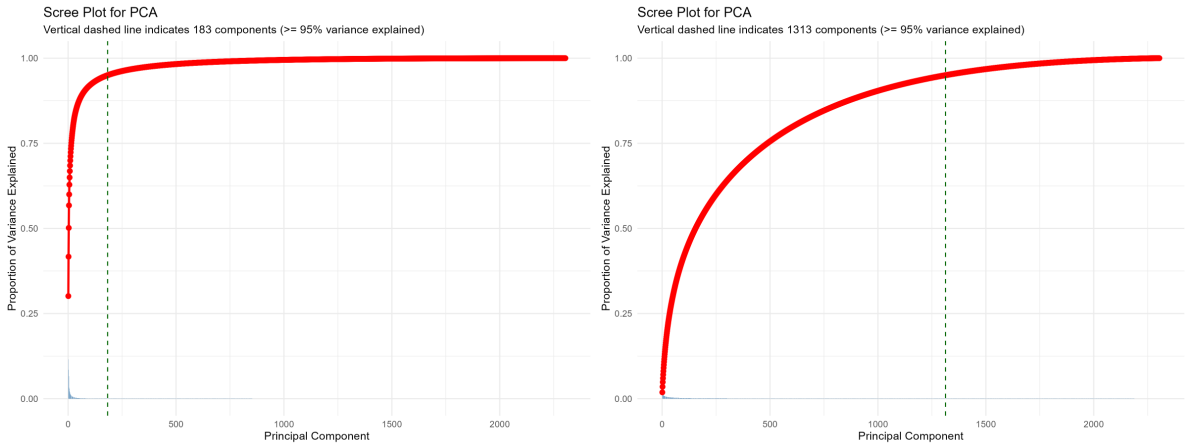

Figure 4: Here are the Combined Scree Plots. Notice how feature engineering transforms the features such that more Principal Components are needed to represent the data.

In our data pipeline, we observed distinct strengths across different base models. XGBoost, when trained without pre-processing but with PCA applied, achieved the highest overall accuracy, effectively balancing precision and recall by capturing the most critical data patterns. The Support Vector Machine (SVM) model, without any pre-processing or PCA, performed well by naturally handling the high-dimensional input space, leveraging its capacity to find optimal decision boundaries even with raw features. Meanwhile, the Random Forest model, trained with both pre-processing and PCA, excelled at minimizing false negatives, making it particularly effective for ensuring that instances of drowsiness were rarely missed. Each model contributed unique advantages, highlighting the importance of diverse strategies within our ensemble learning approach. Each had a relatively high accuracy so these models were chosen for our stacked model:

model	Accuracy	Precision	Recall	FOne	AUC
XGBoostNoFeatureEngineeringPCA	0.8623482	0.9966102	0.7443038	0.8521739	0.9534572
SVMNoFeatureEngineering	0.8596491	1.0000000	0.7367089	0.8483965	0.9852345
SVMNoFeatureEngineeringPCA	0.8421053	1.0000000	0.7037975	0.8261516	0.9632399
RandomForestWithFeatureEngineering	0.8002699	0.7787810	0.8734177	0.8233890	0.9051182
XGBoostNoFeatureEngineering	0.7692308	0.9955752	0.5696203	0.7246377	0.9771713
SVMWithFeatureEngineering	0.7597841	0.7705736	0.7822785	0.7763819	0.8133972
RandomForestNoFeatureEngineeringPCA	0.7503374	0.9906542	0.5367089	0.6962233	0.9358345
SVMWithFeatureEngineeringPCA	0.7449393	0.7754011	0.7341772	0.7542263	0.8297066

model	Accuracy	Precision	Recall	FOne	AUC
RandomForestNoFeatureEngineering	0.7246964	0.9948187	0.4860759	0.6530612	0.9695471
XGBoostWithFeatureEngineering	0.7112011	0.7201946	0.7493671	0.7344913	0.7971903
XGBoostWithFeatureEngineeringPCA	0.5775978	0.8416667	0.2556962	0.3922330	0.7360357
RandomForestWithFeatureEngineeringPCA	0.5614035	0.6902174	0.3215190	0.4386874	0.6895039

And their respective AUC ROC plots:

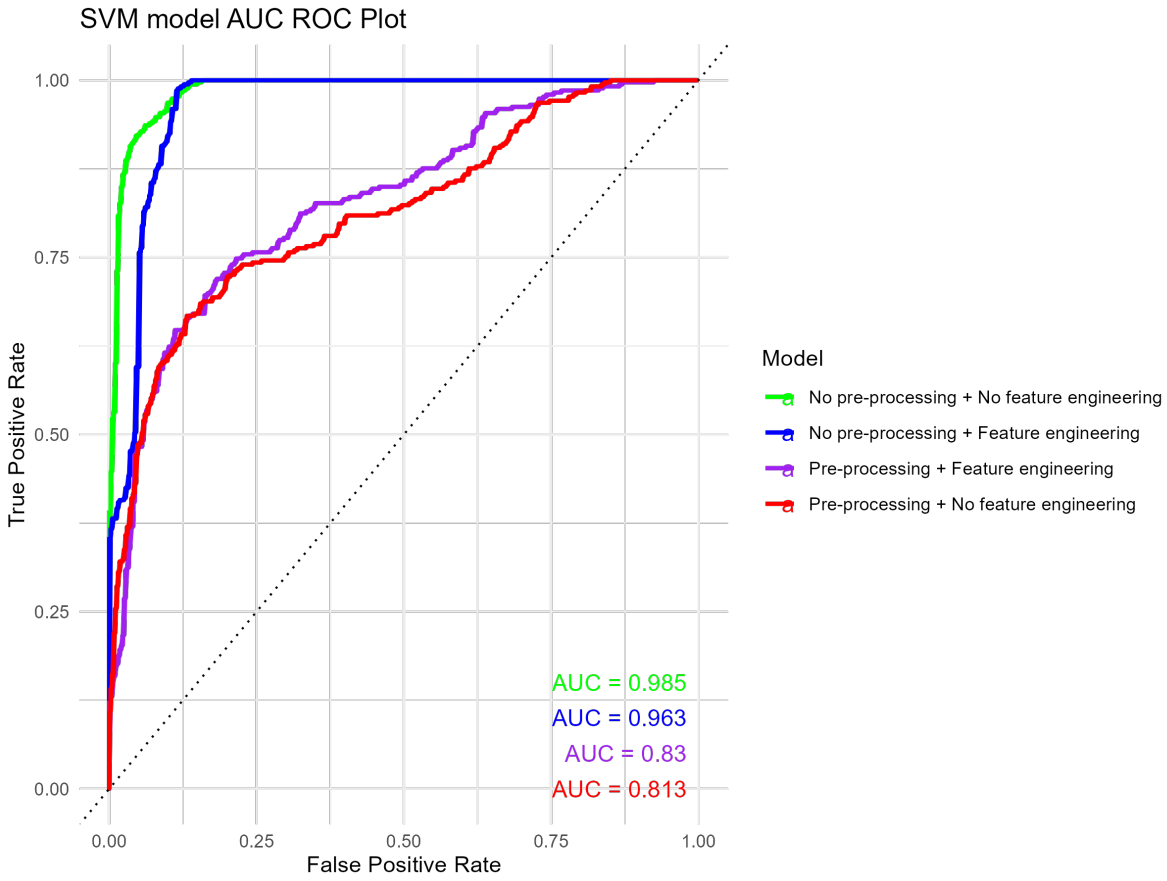

Figure 5: The SVM Base Models. These generally tend to score higher AUC's than the other models.

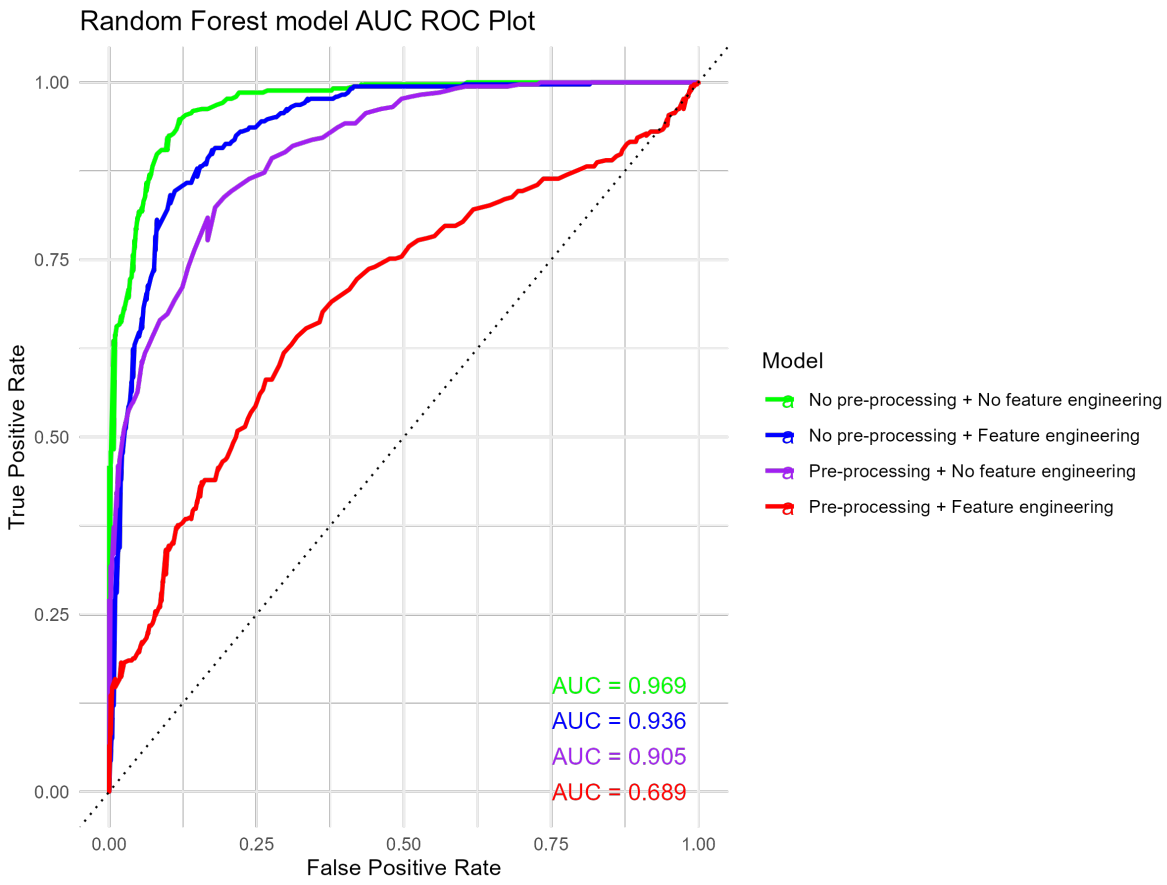

Figure 6: The Random Forest Base models. Notice how for the Random Forest Classifier, the Feature Extraction with PCA has a fairly high AUC.

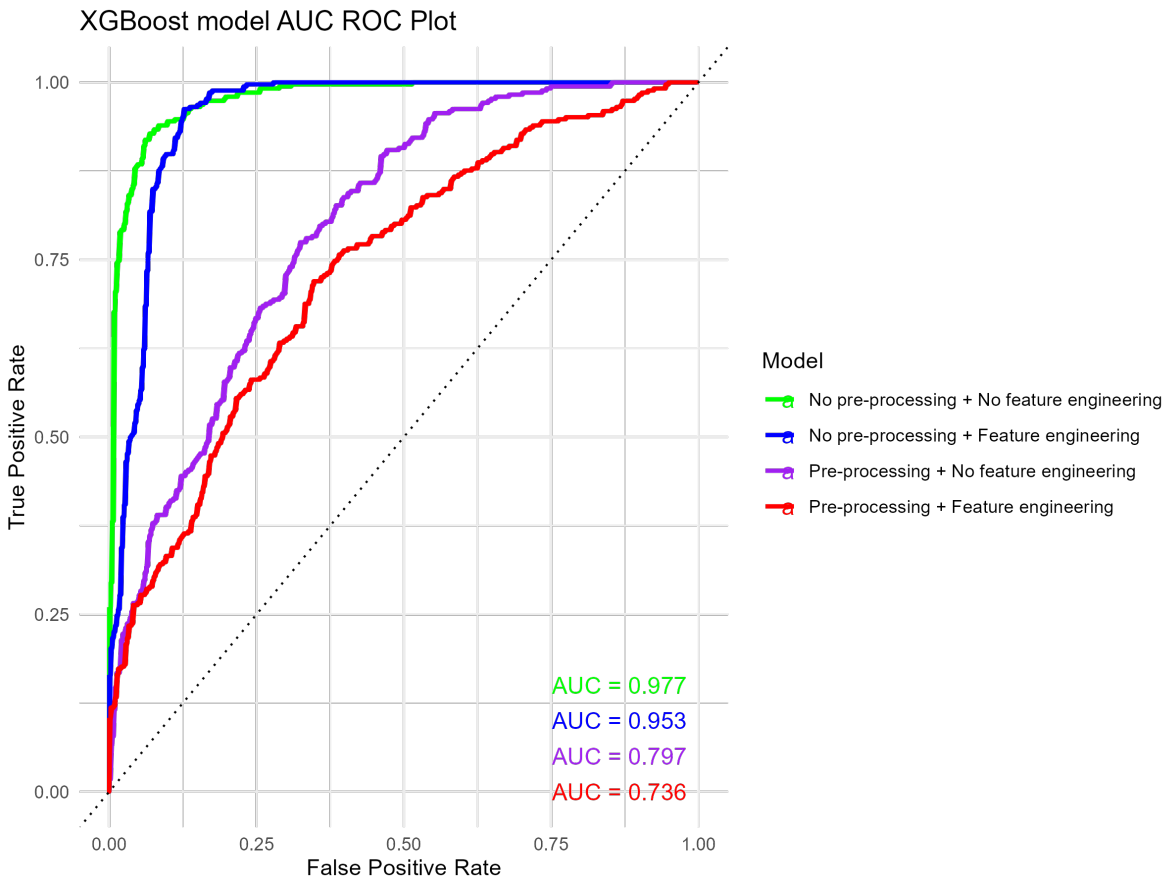

Figure 7: The XGBoost Base models. Notice how it follows roughly the same trend as SVM, but there is a major dip in discriminatory ability when more features are involved.

Something of note is that feature engineering and pre-processing seemed to have obscured most of the models' abilities to discriminate, significantly lowering AUC ROC scores across the board. For those without pre-processing models exhibited a high precision, but a low recall. The biggest anomaly however, was a random forest model that used feature pre-processing and PCA, which had a both a fairly strong precision AND recall. We used empirical information to pick the best 3 models.

In the final stage of our data pipeline, we implemented a Stacking Classifier and meta modeling to further enhance predictive performance. We chose to use a Stacking Classifier because it provides better interpretability while effectively combining the strengths of our top three diverse models. XGBoost contributed by capturing complex non-linear patterns, offering the most informative features for classification. The Support Vector Machine (SVM) complemented this by providing a strong linear decision boundary, while the Random Forest (RF) model, although adding minimal unique information, still contributed robustness to the en-

semble. Importantly, all three models exhibited similar levels of predictability, making them well-suited for stacking. Meta modeling was introduced to refine the ensemble further: it allowed the system to learn an optimal probability fusion plane within the three dimensional prediction space created by the outputs of the stacked models. This approach ensured that the final decision boundary maximized overall classification performance, leading to significant improvements over any individual model. The following results are shown:

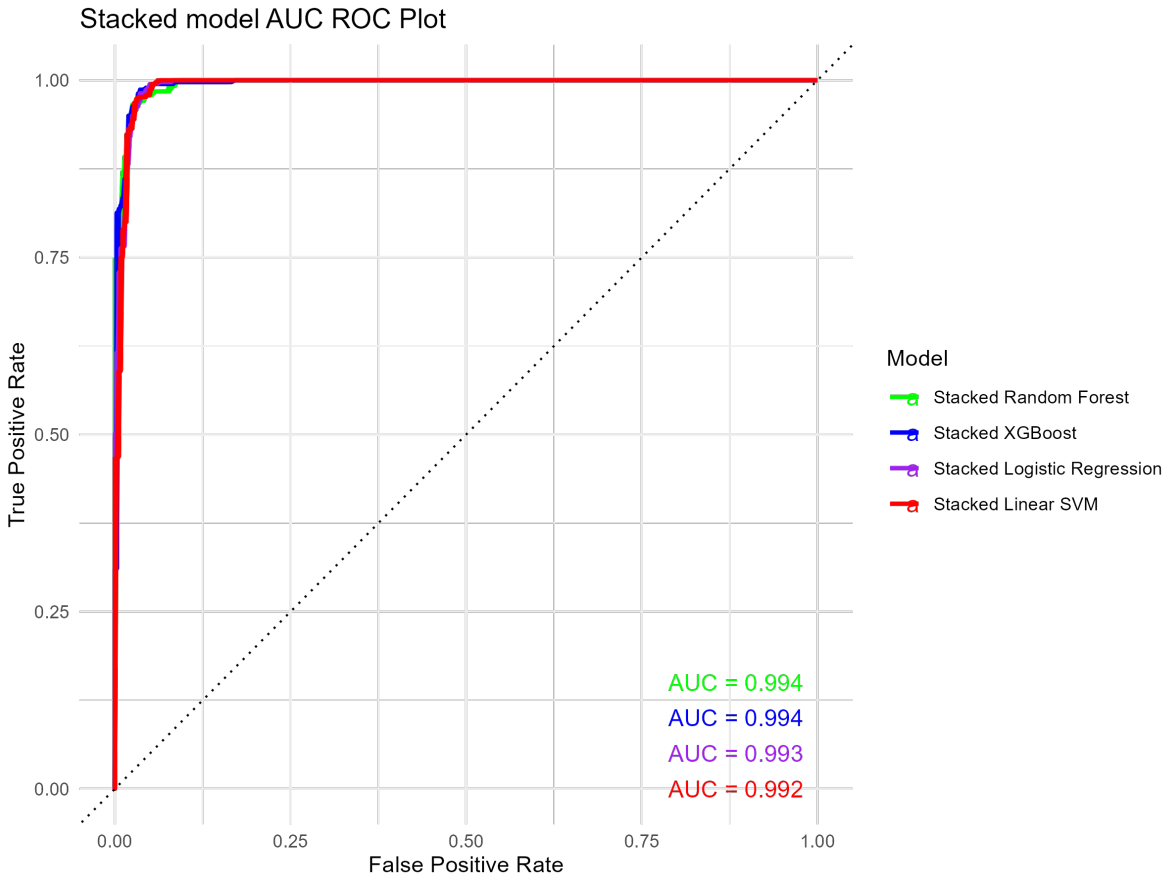

Figure 8: Final Stacked Models

Following their respective classification reports:

model	Accuracy	Precision	Recall	FOne	AUC
XGBoost	0.9690027	0.9643836	0.9723757	0.9683631	0.9942789
LogReg	0.9690027	0.9721448	0.9640884	0.9680999	0.9935010
RandomForest	0.9663073	0.9566396	0.9751381	0.9658003	0.9950894
LinSVM	0.9663073	0.9800570	0.9502762	0.9649369	0.9927523

Given the model results, logistic regression stands out as a strong choice for the drowsiness detection task due to its balance of performance, simplicity, and practicality. Although XGBoost achieved the highest accuracy at 97.0%, logistic regression was nearly identical with 96.9% accuracy and comparable values across precision, recall, F1-score, and AUC. This marginal difference does not justify the added complexity and computational demands of XGBoost, especially in a real-time application like driver monitoring. Logistic regression is much simpler and more interpretable, making it easier to explain and justify decisions to stakeholders in safety-critical environments. It also requires significantly less computational power, which is advantageous for deployment on edge devices such as in-vehicle monitoring systems. Moreover, logistic regression demonstrated consistent and balanced performance across all evaluation metrics, effectively minimizing both false positives and false negatives. This reliability, combined with its ease of implementation and interpretability, makes logistic regression a practical and responsible choice for detecting drowsiness in drivers.

Conclusion:

In this project, we developed a robust and highly accurate classifier for detecting driver drowsiness using facial images. We conducted extensive pre-processing, including automatic CLAHE and dual gamma correction, and optimized feature extraction through Principal Component Analysis (PCA). We trained and evaluated multiple models including XGBoost, Random Forest, Support Vector Machine (SVM), and Logistic Regression using Bayesian hyperparameter tuning and 5-fold cross-validation. Ultimately, we selected Logistic Regression as our final model due to its simplicity, strong interpretability, and near equal performance compared to more complex models like XGBoost. This classifier could be integrated into real world driver monitoring systems, providing an early warning system to reduce accidents caused by drowsy driving. However, there are important areas for future work. One improvement would be to rigorously test all models on entirely separate holdout datasets to further ensure generalizability and prevent any hidden data leakage, which can artificially inflate performance if preprocessing steps unintentionally leak test set information.

Future Works

Future versions of this project could explore integration with existing vehicle technologies like Lincoln's Lane Departure Warning systems, combining drowsiness detection with lane monitoring for enhanced driver safety. From an operational standpoint, implementing ML Ops practices would be critical automating model retraining, monitoring real-world performance drift, and ensuring that the system continues to perform reliably after deployment. Overall, while our current model is robust and not entirely ready for practical use, careful attention to deployment practices, continuous validation, and integration with broader vehicle systems would maximize its impact.

Secondly, because this machine learning classifier requires flattening of the dataset, spatial information is generally lost and not able to be applied to help in the classification task. Alternatives such as Convolutional Neural Networks (CNNs) may be a stronger solution that is able to capture spatial feature information that is lost in our pre-processing pipeline. Both methods can help provide a valuable product that mitigates the dangers of drowsy driving, and can be considered as future expansions of the project. Overall, our study establishes a practical, extensible framework for real-time drowsiness detection that, with continued refinement and integration of advanced spatial models, has the potential to significantly improve road safety and save lives, and with these possible expansions, work in this area of study can be compounded on.

References

[1] Numan Ahmad, Behram Wali, and Asad J. Khattak. "Heterogeneous ensemble learning for enhanced crash forecasts – A frequentist and machine learning based stacking framework". In: *Journal of Safety Research* 84 (2023), pp. 418–434. ISSN: 0022-4375. DOI: https://doi.org/10.1016/j.jsr.2022.12.005. URL: https://www.sciencedirect.com/science/article/pii/S002243752200202X.

[2] Yakun Chang et al. "Automatic Contrast-Limited Adaptive Histogram Equalization With Dual Gamma Correction". In: *IEEE Access* 6 (2018), pp. 11782–11792. DOI: 10.1109/ACCESS.2018.2797872.

[3] Saso DZeroski and Bernard Zenko. "Is Combining Classifiers with Stacking Better than Selecting the Best One?" In: *Machine Learning* 54.3 (Mar. 2004), pp. 255–273. ISSN: 1573-0565. DOI: 10.1023/B:MACH.0000015881.36452.6e. URL: https://doi.org/10.1023/B:MACH.0000015881.36452.6e.

[4] Yashar Jebraeily, Yousef Sharafi, and Mohammad Teshnehlab. "Driver Drowsiness Detection Based on Convolutional Neural Network Architecture Optimization Using Genetic Algorithm". In: *IEEE Access* 12 (2024), pp. 45709–45726. URL: https://api.semanticscholar.org/CorpusID:268761612.

[5] Maine Department of Transportation. *Sleep Deprived? How Does this Compare to Drinking too Much Alcohol?* Tech. rep. Maine Department of Transportation, Nov. 2021. URL: https://www.maine.gov/mdot/challengeme/topics/docs/2021/1121SleepDeprivation.pdf.

[6] Brian C. Tefft. *Prevalence of Motor Vehicle Crashes Involving Drowsy Drivers, United States, 2009–2013.* Tech. rep. 607 14th Street, NW, Suite 201, Washington, DC 20005: AAA Foundation for Traffic Safety, Nov. 2014. URL: https://newsroom.aaa.com/wp-content/uploads/2019/06/AAAFoundation-DrowsyDriving-Nov2014.pdf.