Driver Activity Recognition with Imbalanced Data

Henry Cai
st166080@stud.uni-stuttgart.de

Nikhil Bhavikatti
st188468@stud.uni-stuttgart.de

Hananeh Shizadnia
st190837@stud.uni-stuttgart.de

Remziye Maral Demirsecen
st191631@stud.uni-stuttgart.de

Abstract

Recognizing driver activities, particularly high-risk behaviors, is critical for enhancing road safety. However, real-world datasets often suffer from severe class imbalance, where risky activities are significantly underrepresented. This paper investigates strategies to address class imbalance, including loss function modifications, weighted sampling, and synthetic data augmentation. We conduct experiments across full model training, mean per-class accuracy, and classifier-only training configurations. Our results show that focal loss and normal training strategies achieve strong performance, while weighted sampling and SMOTE, though not outperforming in overall accuracy, demonstrate a focus on improving recognition of minority classes. These findings highlight the importance of balanced training strategies and feature extractor updates for effective driver activity recognition in imbalanced datasets. The code of the project is available at `https://github.com/nikhilbhavikatti/DAR-Imbalanced-Data`

1 Introduction

Based on historical accident statistics, risky driver behavior has been identified as a major contributor to traffic accidents. Risky activities encompass violations of traffic regulations and irregular driving behaviors, making it essential to understand their frequency to develop effective accident prevention strategies [1]. Research indicates that secondary activities behind the steering wheel contribute significantly to traffic accidents, with an estimated 36% of crashes potentially preventable by eliminating such distractions [2]. Therefore, mitigating these risk factors is crucial for reducing the severity of road accidents and associated injuries.

Driver activity data is often imbalanced, where instances of risky driving behaviors occur far less frequently than normal driving activities in real-world traffic scenarios. Supervised learning algorithms trained on imbalanced datasets tend to exhibit poor predictive performance on the minority class, as they inherently prioritize the majority class - normal driving behaviors. However, the primary objective in this context is often to accurately recognize and classify instances of risky driving [1].

To address this challenge, this paper explores multiple strategies for handling class imbalance. First, cross-entropy loss is employed to measure the discrepancy between predicted probabilities and true class labels, serving as a foundational loss function in classification tasks. While cross-entropy loss does not inherently address class imbalance, it provides a baseline for performance evaluation. To mitigate the issue, focal loss is introduced, which down-weights well-classified samples, thereby shifting the model's focus toward harder-to-classify minority class instances. Additionally, weighted sampling techniques, including weighted cross-entropy and weighted focal loss, are implemented to adjust class distributions by oversampling underrepresented classes and undersampling dominant ones during training. Finally, the Synthetic Minority Over-sampling Technique (SMOTE) is applied

to generate synthetic instances of the minority class, enhancing dataset balance and improving the classifier's ability to generalize to underrepresented categories.

2 Related Work

Several studies have addressed the challenge of imbalanced datasets using cost-sensitive learning, which increases the classification error cost for the minority class to counteract the bias introduced by an imbalanced class distribution. Alternatively, ensemble learning algorithms that integrate sampling with bagging or boosting, such as SMOTEBoost, RUSBoost, and EasyEnsemble, have demonstrated superior performance compared to traditional ensemble methods without sampling. Comparisons between cost-sensitive learning techniques - including instance weighting and threshold adjustment - and sampling-based methods have yielded mixed results, as no single approach consistently outperforms the others [1].

Another commonly used technique is under-sampling, where instances from the majority class are removed to balance the dataset. However, this approach has a significant drawback: it reduces the overall dataset size, potentially discarding valuable information. As a result, under-sampling is often less effective when the class imbalance is particularly severe [3].

Hybrid approaches that combine sampling and cost-sensitive learning have also been explored to mitigate class imbalance. However, these strategies have not been extensively applied to driver activity recognition. For instance, Le et al. [4] integrated sampling with a Cluster-based Boosting algorithm within a cost-sensitive learning framework for bankruptcy prediction. Similarly, Peng et al. [5] employed a combination of sampling and cost-sensitive multilayer perceptrons (MLPs) for traffic accident prediction. Despite their effectiveness, these methods rely on predefined sampling strategies, fixed sampling ratios, and pre-determined weight assignments in cost-sensitive classifiers. This rigidity can lead to suboptimal solutions and, in some cases, may result in worse performance compared to using either sampling or cost-sensitive learning alone [1].

The challenge of learning from imbalanced datasets can also be framed as a label shift problem in transfer learning and domain adaptation. In such scenarios, the primary difficulty lies in detecting and estimating label shift. Once label shift is identified, techniques such as re-weighting or re-sampling are applied to adjust for distribution discrepancies [6].

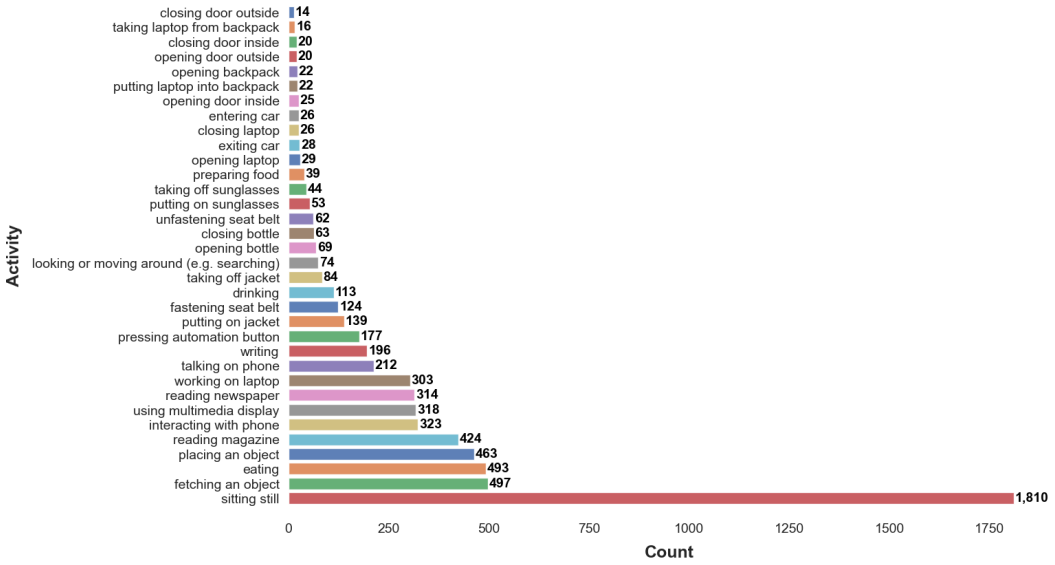


Figure 1: Class distribution of activities in the Drive&Act dataset. The dataset is highly imbalanced, with certain activities occurring significantly more frequently than others.

3 Dataset

The publicly released Drive&Act dataset is utilized to address driver activity recognition under imbalanced data conditions. This dataset consists of over 12 hours of recordings and more than 9.6 million frames, capturing individuals engaged in various distracting activities during both manual and automated driving scenarios. It includes RGB, infrared (NIR), depth, and 3D body pose information, recorded from six different viewpoints. A hierarchical annotation system is applied, providing dense labeling across 83 activity categories [2]. For our experiments, we specifically used a subset of Drive&Act: the A-Column co-driver NIR data along with its corresponding activity annotations.

Within this subset, significant class imbalance is observed in activity distributions. Figure 1 illustrates the frequency of different driver activities. Some activities, such as "sitting still" (1,810 instances), dominate the dataset, whereas others, such as "closing door outside" (14 instances), are heavily underrepresented. This imbalance presents challenges for supervised learning models, as they tend to be biased towards majority classes. Addressing this issue requires specialized techniques such as re-weighting loss functions, sampling strategies, and synthetic data augmentation to ensure fairer model performance across all activity categories.

4 Model Architecture

The proposed model for driver activity recognition is based on MobileNetV2 3D Jester, a computationally efficient deep neural network designed for resource-constrained environments. MobileNetV2 3D Jester employs depth-wise separable convolutions to significantly reduce computational complexity while maintaining high feature extraction capabilities. A key innovation in this architecture is the inverted residual block with linear bottlenecks, which enhances information flow by using narrow intermediate layers for efficient representation learning. These bottleneck layers consist of an expansion phase, where the feature dimension is increased, followed by depth-wise convolution and a projection phase that maps features back to a lower-dimensional space. This structure enables improved gradient propagation and reduces memory overhead. The network concludes with a global average pooling layer, followed by a fully connected classification layer that outputs the predicted driver activity class. Given its lightweight nature and strong feature extraction capabilities, MobileNetV2 3D Jester is well-suited for real-time activity recognition tasks, particularly when working with imbalanced datasets, where efficient representation learning is crucial for capturing minority class patterns.

5 Methodology

5.1 Loss Functions

5.1.1 Cross Entropy Loss

Cross-Entropy Loss (CE) is a widely used loss function in machine learning, particularly for classification tasks. It measures the difference between the true label distribution and the predicted probability distribution. The basic idea behind cross-entropy is to penalize the model for deviating from the true labels. Formally, for a classification problem, the cross-entropy loss is computed as:

$$L_{CE} = -\sum_{i} y_{i,n} \log(z_{i,n})$$

where $y_{i,n}$ represents the true label and $z_{i,n}$ represents the predicted probability for the n-th class. The function assigns a higher penalty when the model predicts a low probability for the correct class, which helps the model adjust its parameters during training.

In our work, cross-entropy loss evaluates model performance in driver activity recognition by measuring the difference between true and predicted labels. It minimizes error by assigning higher probabilities to correct activities, improving classification accuracy. However, it may struggle with class imbalance, where distracted driving behaviors are less frequent. To address this, alternative loss functions like focal loss enhance minority class recognition.

5.1.2 Focal Loss

Focal Loss (FL) is a loss function designed to address the challenge of class imbalance in classification problems. It is particularly effective in scenarios where the distribution of classes is highly imbalanced, and the model struggles to accurately classify the less frequent, harder-to-predict classes. FL is a variant of Weighted Cross Entropy (WCE) and works by dynamically adjusting the loss function based on the prediction accuracy. Specifically, the loss is weighted by a factor that depends on the error between the predicted probability $z_{i,n}$ and the true label $y_{i,n}$. The weighting factor is given by:

$$L_n^{FL} = \alpha \sum_i (1 - z_{i,n})^\gamma y_{i,n} \log(z_{i,n})$$

In this formulation, examples that are difficult to classify - i.e., those with lower prediction accuracy - contribute more to the total loss, while easier-to-classify examples have a reduced impact. The term $(1 - z_{i,n})^\gamma$ increases the loss for misclassified examples, especially when $z_{i,n}$ is close to zero. The coefficients α and γ, which are consistent across all classes, control the strength of the loss function, with γ governing the importance of hard-to-classify examples [7].

Focal Loss is particularly useful in applications such as driver activity recognition, where there is an inherent class imbalance in the dataset. In these cases, the model may face difficulty in accurately classifying the less frequent, harder-to-detect distracted driving activities. By applying Focal Loss, the model dynamically adjusts the weighting of the loss function and improves the model's ability to focus on these harder-to-classify instances.

5.2 Data-Level Methods

5.2.1 Weighted Sampling

Weighted sampling is a technique that adjusts the training process by assigning higher weights to minority classes during batch sampling. This ensures that the model sees more examples from underrepresented classes, thereby improving its ability to learn rare activities. The weight for each class is typically set inversely proportional to its frequency in the dataset:

$$w_c = \frac{N}{C \cdot n_c}$$

where w_c is the weight for class c, N is the total number of samples, C is the number of classes, and n_c is the number of samples in class c.

In our implementation, we used weighted sampling in conjunction with cross-entropy loss and focal loss to balance the dataset.

5.2.2 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a re-sampling method that generates synthetic examples for minority classes by interpolating between existing samples. Unlike random over-sampling, which simply duplicates minority class examples, SMOTE creates new, synthetic examples that lie along the line segments connecting k-nearest neighbors in the feature space. This helps balance the dataset and improves the model's ability to learn minority classes[8].

The SMOTE algorithm works as follows:

1. For each minority class sample x_i, find its k-nearest neighbors.
2. Randomly select one of the k neighbors, x_j.
3. Generate a synthetic sample x_{new} by interpolating between x_i and x_j:

$$x_{\text{new}} = x_i + \lambda(x_j - x_i)$$

   where λ is a random number between 0 and 1.

In our experiments, SMOTE was applied to the training dataset to balance the class distribution.

6 Experiments

This section presents the experimental setup designed to evaluate the performance of the proposed model under various training configurations. Four distinct experiments were conducted, each addressing different aspects of model learning and generalization. The experiments varied in terms of the training strategy, evaluation metric, and data preprocessing techniques employed.

6.1 Full Model Training

The first experiment involved training the entire model end-to-end, updating both the feature extractor and classifier simultaneously. The objective was to evaluate the model's ability to learn spatiotemporal representations from the input data. The model was trained using two different loss functions (cross entropy loss and focal loss) and two sampling strategies (normal and weighted sampling) across three learning rates: 10^{-2}, 10^{-3}, and 10^{-4}. The training was conducted for 50 epochs, with validation and test accuracy recorded at predefined checkpoints. This experiment serves as the baseline for assessing improvements in the subsequent configurations.

6.2 Training with Mean Per-Class Accuracy

To mitigate potential class imbalance effects, a second experiment was conducted in which model performance was evaluated using mean per-class accuracy rather than overall accuracy. The training procedure remained identical to the full model training experiment; however, the selection of the best-performing model was based on its mean per-class accuracy rather than conventional accuracy. This evaluation strategy ensured that the model did not disproportionately favor majority classes at the expense of underrepresented ones.

6.3 Classifier-Only Training

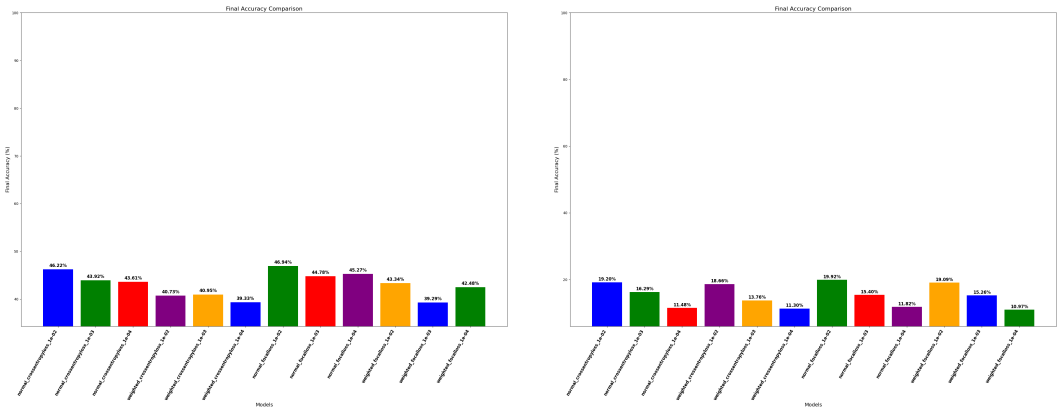
In the third experiment, the feature extractor was frozen, and training was restricted to the classifier. This setup was designed to isolate the contribution of the classifier in distinguishing between activities while leveraging pre-trained feature representations. In addition to normal and weighted sampling, this experiment incorporated synthetic data augmentation through SMOTE to address class imbalance. The training was performed using the same loss functions and learning rates as in the previous experiments, over a total of 1000 epochs.

6.4 Classifier-Only Training with Reduced Majority Class

A fourth experiment was conducted to analyze the impact of reducing the dominance of the majority class. Specifically, 75% of the training instances corresponding to the activity "sitting still" were removed, as this activity was observed to be significantly overrepresented in the dataset. The objective was to determine whether reducing the frequency of the most common class would lead to improved model generalization across other activities. The model was trained using the classifier-only setting, employing the same sampling strategies and loss functions as in the third experiment. The results were compared to those obtained in previous settings to assess the influence of majority-class.

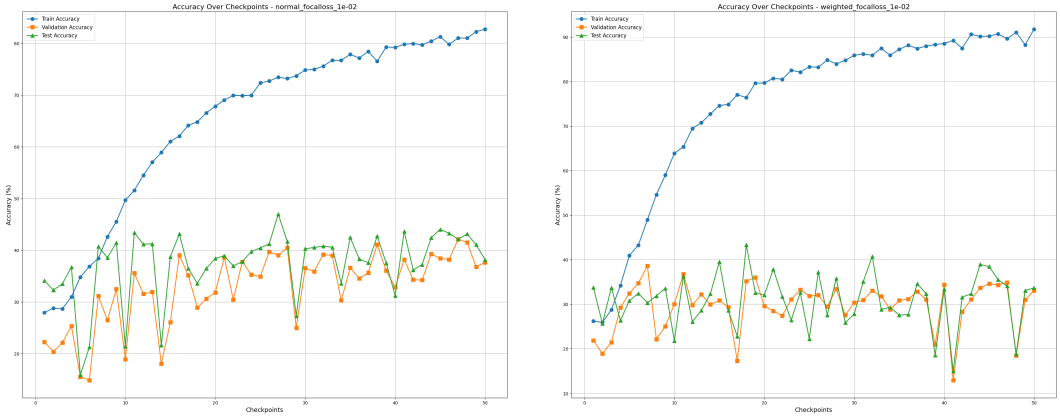
7 Results and Discussion

The experimental results demonstrate the influence of different training strategies on model performance, with notable variations across configurations. In the full model training setup, models trained without weighted sampling consistently outperformed those using weighted sampling. The best-performing model in this category, utilizing focal loss at a learning rate of 10^{-2}, achieved a test accuracy of 46.94%, compared to 43.34% for the best-weighted sampling model under the same conditions. Figure 2a provides a comprehensive comparison of all models in this experiment, while Figure 3 and Figure 4 offer detailed insights into the training progression and label-wise performance of the top-performing normal and weighted models. These results suggest that weighted sampling, while often employed to address class imbalance, may not always yield superior performance in this context. Furthermore, training with mean per-class accuracy, designed to mitigate class imbalance effects, produced comparable results between normal and weighted models, with the best normal



(a) Full Model Training (b) Training with Mean Per-Class Accuracy

Figure 2: Comparison of model test accuracy between baseline and weighted sampling approaches, evaluated using Cross-Entropy Loss and Focal Loss across three learning rates (10^{-2}, 10^{-3}, and 10^{-4}). Results are shown for both full model training and training with mean per-class accuracy.



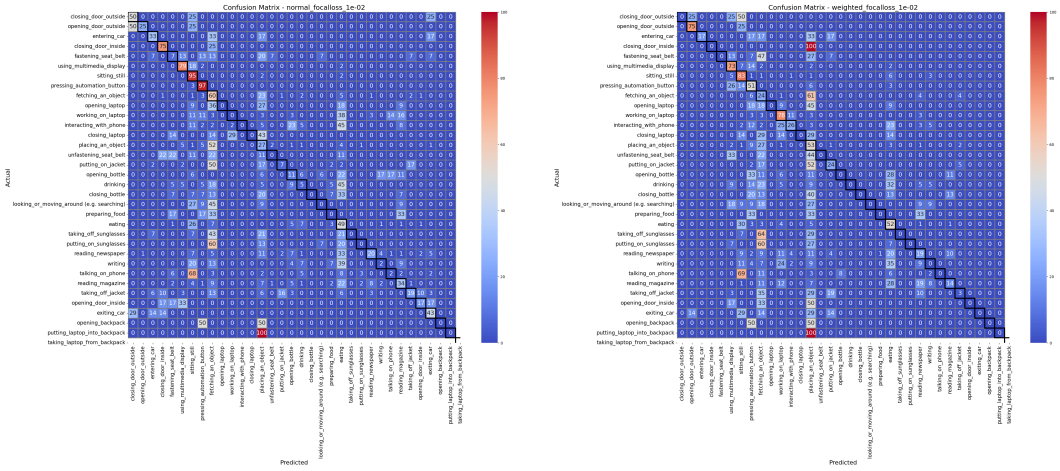
(a) Normal Focal Loss (b) Weighted Focal Loss

Figure 3: Training, validation, and test accuracy across 50 epochs for normal and weighted focal loss, evaluated at a learning rate of 10^{-2} during full model training.

model achieving 19.92% and the best-weighted model achieving 19.09% test accuracy. Figure 2b illustrates the performance distribution across models in this setting, indicating that while mean per-class accuracy ensures balanced class representation, it does not significantly enhance overall accuracy compared to conventional metrics.

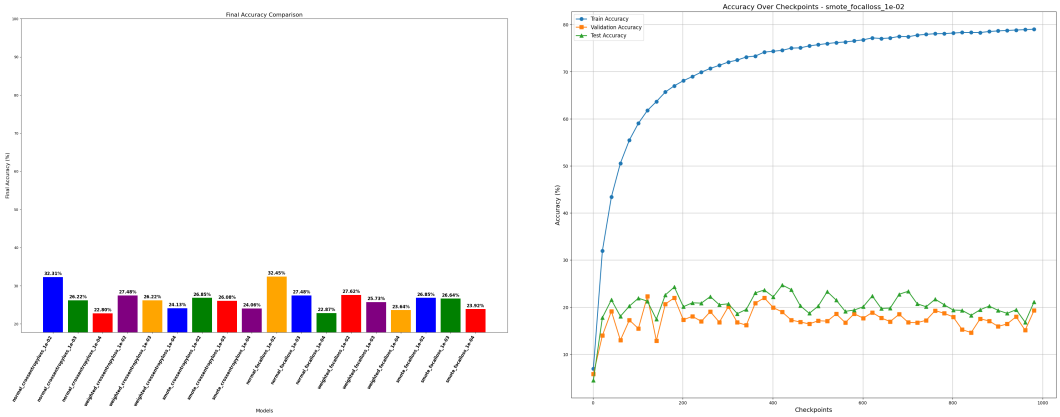
In the classifier-only training experiments, normal models again demonstrated superior performance relative to weighted sampling and SMOTE-augmented models. The best normal model achieved a test accuracy of 32.45%, surpassing the best weighted model (27.62%) and the best SMOTE model (26.85%). Figure 5a provides a comprehensive comparison of all models in this experiment, while Figure 5b and Figure 6 provide additional insights into the training dynamics of the best SMOTE model, trained at a learning rate of 10^{-2}. These findings underscore the critical role of feature extractor updates in achieving optimal model performance. However, the classifier-only training configuration with a reduced majority class resulted in significantly degraded performance, with all models achieving less than 12% accuracy. This outcome highlights the challenges posed by severe class imbalance and the limitations of classifier-only training in such scenarios, emphasizing the need for balanced datasets and holistic training strategies to ensure robust generalization.

Figure 6 presents confusion matrices for SMOTE with focal loss trained at two different learning rates (10^{-2} and 10^{-3}). The comparison highlights key differences in how the learning rate affects minority-class recognition and overall model performance. At 10^{-3}, the model correctly classifies a



(a) Normal Focal Loss (b) Weighted Focal Loss

Figure 4: Normalized Confusion matrices for normal and weighted focal loss, evaluated at a learning rate of 10^{-2} during full model training.

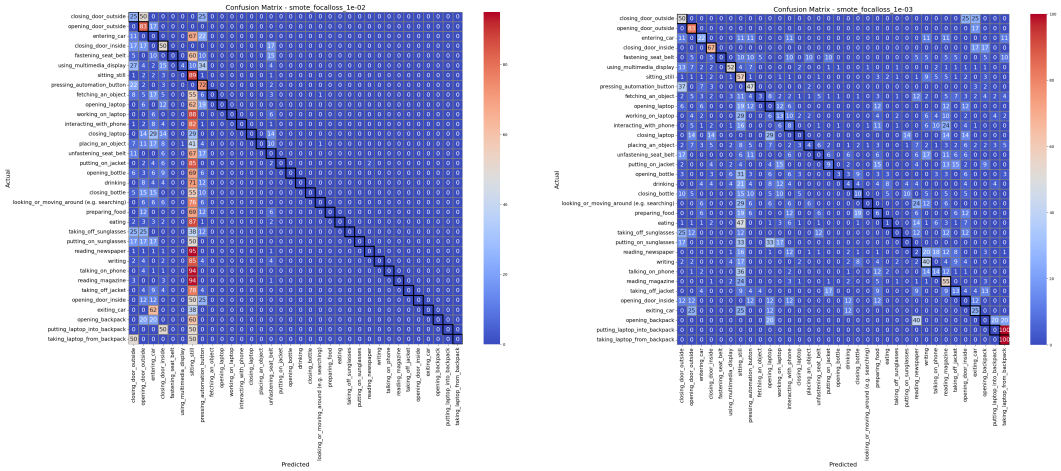


(a) (b)

Figure 5: a) Comparison of model test accuracy between baseline, weighted sampling, and SMOTE approaches, evaluated using Cross-Entropy Loss and Focal Loss across three learning rates (10^{-2}, 10^{-3}, and 10^{-4}) for classifier-only training. b) Training, validation, and test accuracy across 1000 epochs for SMOTE with focal loss, evaluated at a learning rate of 10^{-2} during classifier-only training.

higher number of minority-class instances compared to 10^{-2}, indicating that a lower learning rate allows better fine-tuning of decision boundaries for underrepresented activities. However, while SMOTE improves recognition of underrepresented activities, Figure 6a shows that at 10^{-2}, the model overfits to certain minority classes, causing misclassifications in majority classes. In contrast, Figure 6b demonstrates that 10^{-3} maintains a better balance but at the cost of slightly lower overall accuracy. These results suggest that SMOTE is highly sensitive to learning rate adjustments, and a higher learning rate may cause instability in synthetic sample integration, leading to errors in distinguishing activity classes. Thus, practitioners should carefully tune the learning rate when applying SMOTE to ensure optimal performance. These findings highlight the need for a balanced approach that integrates loss function selection, dataset balancing strategies, and hyperparameter tuning to achieve optimal driver activity recognition in imbalanced datasets.

In order to better evaluate the new approaches, we would like to conduct a significance test. The significance test seeks to understand if the differences between two distributions are meaningful or random. In our situation these distributions are model accuracies and analyze if the performance of one model significantly exceeds that of the other. In order to conduct this test in repeatable



(a) Learning rate of 10^{-2}

(b) Learning rate of 10^{-3}

Figure 6: Normalized Confusion matrices for SMOTE with focal loss, evaluated at a learning rate of 10^{-2} and 10^{-3} during classifier-only training.

manner, we need sufficient amount of validation data and multiple accuracy figures from each model. Unfortunately, in this case, the dataset we have available is not large enough, and this makes it very challenging for us to work. From the limited data, we get highly variable estimates of accuracy, making it difficult to draw strong conclusions. The additional problem stems from spending considerable resources on multiple experiments while there is insufficient data to fully utilize these resources. Therefore, this places constraints on how robust and informative the assessment of model performance can be, making it much more complex and expensive, and leading to either more data or other means to try to solve these issues.

8 Conclusion

In this study, we evaluated the proposed approaches by conducting a statistical significance test. We investigated various methods to address class imbalance in driver activity recognition, including loss functions such as cross-entropy and focal loss, as well as data-level techniques like weighted sampling and SMOTE. Our results demonstrate that focal loss is the most effective approach for handling class imbalance, consistently achieving superior performance across different training configurations. While weighted sampling and SMOTE were designed to improve minority class recognition, they did not consistently enhance overall accuracy and, in some cases, led to reduced performance, potentially due to unstable training or insufficient regularization. These findings highlight the importance of loss function selection and feature extractor updates over data-level augmentation techniques for optimizing model performance in imbalanced datasets. Future research should focus on exploring the theoretical foundations of focal loss's effectiveness, investigating adaptive sampling strategies, and developing hybrid loss functions to further improve minority class recognition in driver activity datasets.

References

[1] K. Wang, Q. Xue, and J. Lu. Risky driver recognition with class imbalance data and automated machine learning framework. *International Journal of Environmental Research and Public Health*, 18, 2021.

[2] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2801–2810, 2019.

[3] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. 2019.

[4] T. Le, M.T. Vo, B. Vo, M.Y. Lee, and S.W. Baik. A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction. *Complexity*, 2019:1–12, 2019.

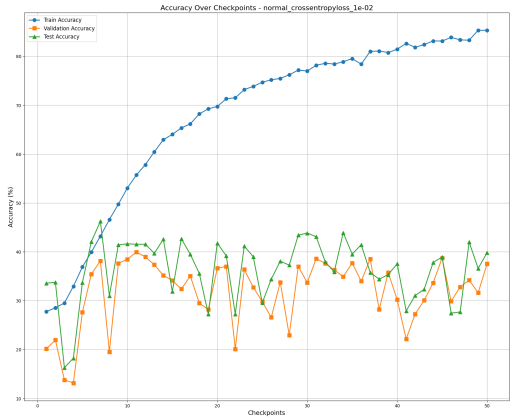
[5] Y. Peng, C. Li, K. Wang, Z. Gao, and R. Yu. Examining imbalanced classification algorithms in predicting real-time traffic crash risk. *Accident Analysis & Prevention*, 144:105610, 2020.

[6] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

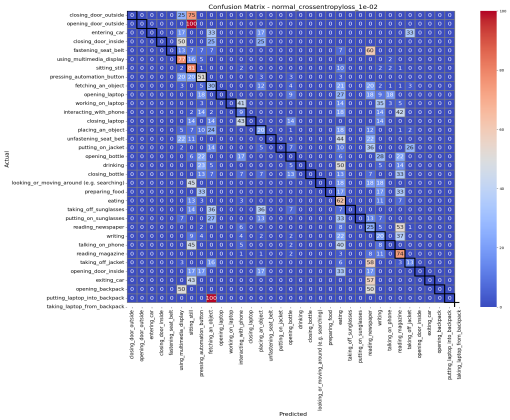
[7] M.S. Hossain, J. Betts, and A. Paplinski. Dual focal loss to address class imbalance in semantic segmentation. *Neurocomputing*, 462, 2021.

[8] L. O. Hall W. P. Kegelmeyer. N. V. Chawla, K. W. Bowyer. Smote: Synthetic minority oversampling technique. *Journal Of Artificial Intelligence Research*, 16, 2011.

A Additional Results



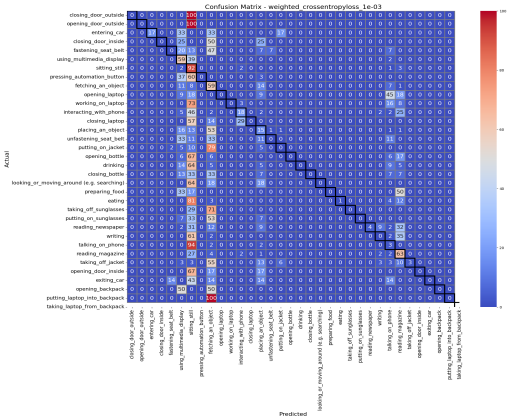
(a) Accuracy over 50 epochs



(b) Normalized Confusion Matrix

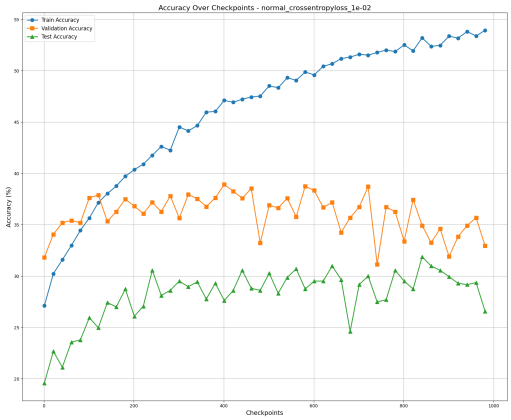
Figure 7: Results for Normal Cross-Entropy Loss evaluated at a learning rate of 10^{-2} during full model training.

(a) Accuracy over 50 epochs

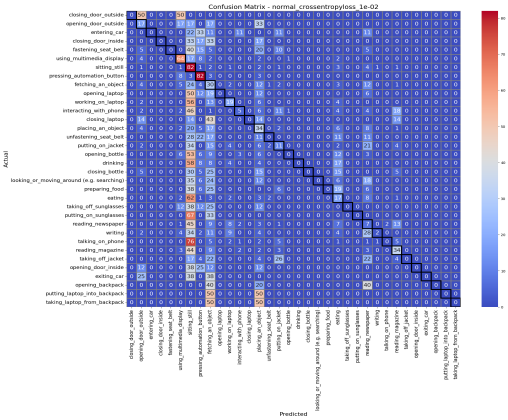


(b) Normalized Confusion Matrix

Figure 8: Results for Weighted Cross-Entropy Loss evaluated at a learning rate of 10^{-3} during full model training.

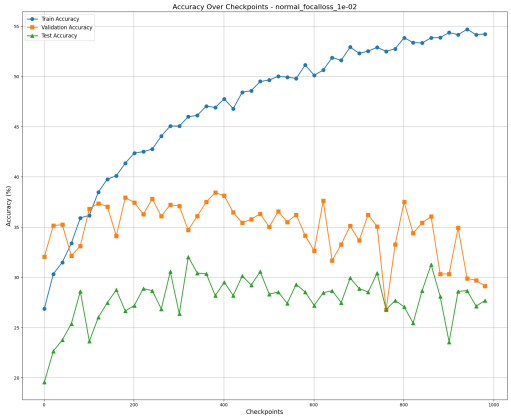


(a) Accuracy over 1000 epochs

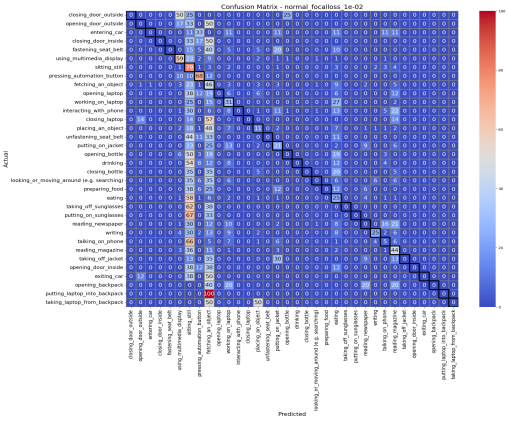


(b) Normalized Confusion Matrix

Figure 9: Results for Normal Cross-Entropy Loss evaluated at a learning rate of 10^{-2} during classifier-only training.



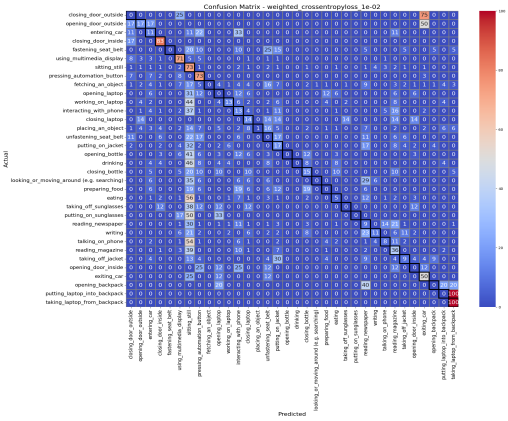
(a) Accuracy over 1000 epochs



(b) Normalized Confusion Matrix

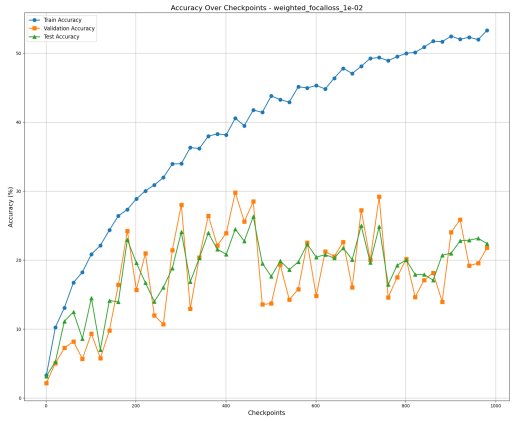
Figure 10: Results for Normal Focal Loss evaluated at a learning rate of 10^{-2} during classifier-only training.

(a) Accuracy over 1000 epochs

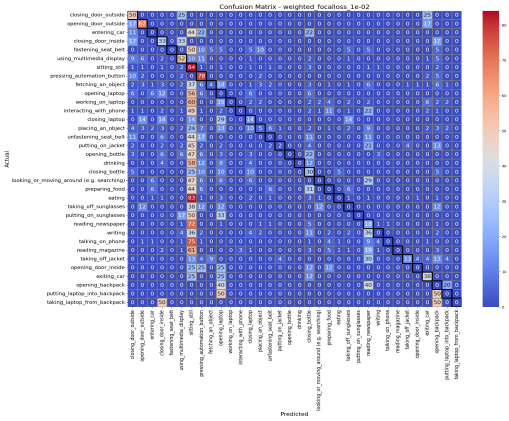


(b) Normalized Confusion Matrix

Figure 11: Results for Weighted Cross-Entropy Loss evaluated at a learning rate of 10^{-2} during classifier-only training.

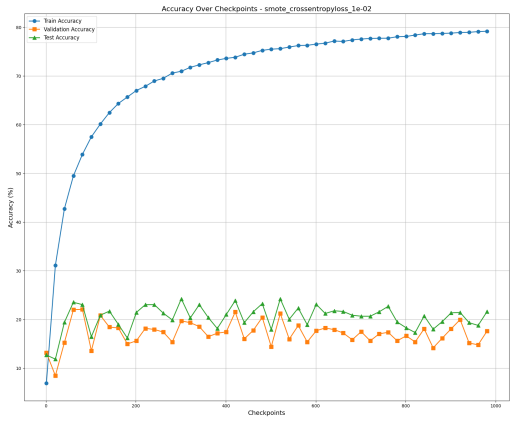


(a) Accuracy over 1000 epochs

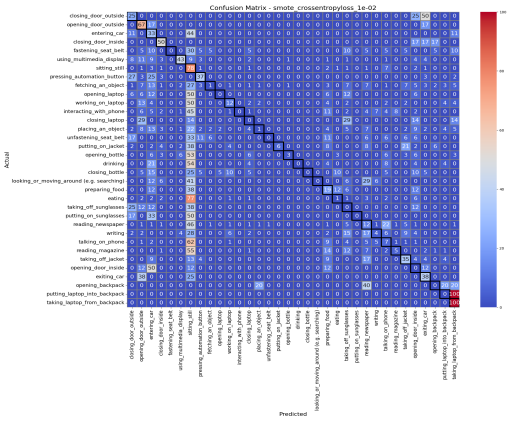


(b) Normalized Confusion Matrix

Figure 12: Results for Weighted Focal Loss evaluated at a learning rate of 10^{-2} during classifier-only training.



(a) Accuracy over 1000 epochs



(b) Normalized Confusion Matrix

Figure 13: Results for SMOTE Cross-Entropy Loss evaluated at a learning rate of 10^{-2} during classifier-only training.