AUTOMATIC CLASSIFICATION OF FOUL EVENTS IN FOOTBALL USING DEEP LEARNING

DAAN JAN GROTENHUIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COGNITIVE SCIENCE AND ARTIFICIAL INTELLIGENCE

AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES OF TILBURG UNIVERSITY

STUDENT NUMBER

259935

COMMITTEE

Drs. Stijn Rotman

Ph.D. Mariana Dias Da Silva-van Riel

LOCATION

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

DATE

June 23rd, 2023

WORD COUNT

6409

ACKNOWLEDGMENTS

I would like to thank my supervisor, Drs. Stijn Rotman, for helping me during this bachelor thesis project. I would also like to thank Jan Held for his help and for allowing me to use and access his dataset and code early before it was public.

This thesis does not suggest the replacement of VAR but merely an addition to the current VAR system based on the limitations mentioned above. This thesis believes the combined effort and supervision of both machines and humans will result in the best results.

AUTOMATIC CLASSIFICATION OF FOUL EVENTS IN FOOTBALL USING DEEP LEARNING

Daan Jan Grotenhuis

**Abstract**

The pursuit of fairness and accurate decision-making by football referees remains a concern within football. Despite the implementation of various systems over the years, human limitations have continued to introduce errors. To address this issue, this thesis investigates automatic classification of foul events in football using deep learning. This thesis leverages the recently released Soccernet-MVFoul dataset. In addition to examining the automatic classification of foul events, this thesis explores two supplementary aspects: the effectiveness of data augmentation and the utilization of unexplored models with this dataset. The design of this thesis incorporates the use of single-view data instead of multi-view data, as well as excludes the use of labels to mitigate complexities, class imbalance, and increase the available data. The thesis adopts widely used models, loss functions, optimizers, and other established methodologies employed in previous works. Results demonstrated an average weighted accuracy of 47% between all models and that data augmentation did not yield significant improvements. This reinforces future endeavors in deep learning being able to further enhance the pursuit of fairness in football.

**DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT**

Data Source: The SoccerNet-MVFouls dataset has been acquired from the author through an online request. The obtained data is anonymized. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. However, the institution was informed about the use of this data for this thesis and potential research publications. All the figures belong to the author. The thesis code can be accessed through the google drive share link (https://drive.google.com/drive/folders/1qo44UrqHxlB1D42XnOHtpqFQr2oTF2AH?usp=sharing). Part of the code has been adapted by the author from (https://drive.google.com/file/d/12SOSeetlL6SjFG9oSvYr6w3YEvM6kHZV/view?usp=sharing). The reused/ adapted code fragments are clearly indicated in the notebook. In terms of writing, the author used assistance with the language of the paper. A generative language model Grammarly (https://app.grammarly.com/) and ChatGPT (https://chat.openai.com/) was used to improve the author’s original content, for paraphrasing, spell-checking, and grammar. No other typesetting tools or services were used.

**1 Introduction**

The football industry is a tremendously popular and profitable sector in the sports industry, being able to generate a revenue of 5.5 billion in the 20/21 season from the top 5 leagues (Premier League, Bundesliga, Serie A, La Liga, Ligue Un) alone (Statista, 2023). A key concern within this industry is ensuring the fairness and accuracy of refereeing. Considering the stakes involved in this sector, minor errors, which may be attributed to human error, in the referee’s judgment and interpretation of fouls may potentially affect the outcome of matches. Furthermore, it could influence the social and financial consequences for the players, fans, referees, and clubs involved.

Regardless of their best efforts, the interpretation of fouls remains vulnerable to inconsistencies and errors by referees. The introduction of the Video Assistant Referee (VAR) system in 2017-19 in the top 5 leagues aimed to combat this issue by providing an off-pitch referee with the ability to review potential infractions and alter incorrect calls made by the on-pitch referee. However, this system is still hindered by differing interpretations of rules by the main and VAR referees, aswell as human error.

A recent example of the impact that incorrect decisions can have on the results of a match involves the recent Arsenal-Brentford match in the Premier League. In this incident, a foul committed on an Arsenal player was not deemed severe enough by the main referee to be called a foul. This consequently led to the challenge not getting reviewed by the VAR referee. Had the foul been correctly recognized and penalized, it would’ve significantly altered the match results. After the error by the main referee came to light, he came forward, publicly acknowledged the mistake, and relieved his duties as a Premier League referee (I. Mustapha, 2023).

This incident serves as a reminder of the continued need for development, despite the recent introduction of systems like the VAR. This thesis project proposes the development of a deep learning-fueled system to further aid the refereeing process and propel the fairness and accuracy of match results. The introduction of such a system could assist the VAR referee, notifying them when a potential violation has been detected and classified, aswell allowing VAR referees to manually review the incident and alert the main referee if necessary.

The main objective of this thesis is to explore the idea of:

*automatic classification of foul events in football using deep learning.*

To address this research objective, this thesis will utilize the recently released Soccer-MVFoul dataset, which serves as a subsidiary of the more commonly used Soccernet databases. Unlike its parent dataset, the Soccernet-MVFoul database is highly specialized in fouls, consisting exclusively of video clips depicting foul events. This narrow focus of the database makes it particularly well-suited to attempt to answer the task at hand. The paper accompanying the release of the dataset from Held et al. (2023) delves into the possibility of an automated VAR for leagues that do not have the capacities or finances for such a system. While our research aligns with their overarching objective, a major distinction between Held's paper and this thesis is the incorporation of data augmentation. This technique is commonly used to combat class imbalance issues in deep learning tasks. However, this thesis hypothesizes that the effectiveness of data augmentation may not always be beneficial for all tasks. Thus, this thesis aims to investigate a sub-research question, which investigates:

*to what degree does data augmentation affect a model’s performance?*

Furthermore, prior research on similar topics as this thesis made use of a wide variety of machine-learning models, as different models have different properties which allow them to perform better under different tasks. Therefore, this thesis will introduce a new model to the Soccernet-MVFoul dataset that was not been explored yet. Doing so possibly allows for improved performance compared to existing models explored in prior research.

Consequently, this thesis aims to investigate an additional sub-research question:

*to what extent does the introduction of a new model increase results?*

**2 Related Work**

There is an abundance of previous work conducted on video event classification using AI. Specialized studies on foul events in football are not as frequent, as numerous event classification research studies that incorporate foul events into general event classification. The study conducted by Vidal-Codina et al. (2022) focuses on the utilization of tracking data for automatic event detection in football. However, the paper treats foul events as a general category, combining various events, such as offside situations into foul events. Similarly, Sen et al. (2022) use a different approach in their investigation, employing transfer learning from multiple models, including resnet, in combination with Gated Recurrent Units (GRU) to classify actions in soccer videos. Encouragingly, their research demonstrates promising results. Similar to Vidal-Codina et al. (2022), Sen et al. (2022) do not distinguish fouls as distinct events within their classification framework.

Nevertheless, it is worth noting that the studies, including Vidal-Codina et al. (2022) and Sen et al. (2022), did not specifically address the differentiation of various foul events such as fouls, yellow cards, and red cards. In contrast, other research, such as Sangram Singh Rana et al. (2023) and Giancola et al. (2021), recognized the significance of differentiating between these specific types of fouls. Utilizing the Soccernet v2 database, the differentiation of foul events allowed for a more precise and accurate classification of foul events in football.

Thamaraimanalan et al. (2020) conducted research specifically focused on the classification of fouls, underlining the significance of differentiating between various types of fouls. It is crucial to recognize that fouls consist of distinct events, and treating them as a single category disregards their differences. For instance, a foul without a booking (no card being given) is different from a foul that results in a booking (yellow or red card given). The severity of the penalty and the impact on the football game vary tremendously between these scenarios. Additionally, missing the detection of a regular foul event might have a minor impact compared to overlooking a red card event, which carries significantly higher consequences and implications for the match outcome. Therefore, accurately classifying and differentiating foul events is of substantial importance in football research and analysis.

Numerous noteworthy papers, as listed above, on foul detection or general event detection utilize the Soccernet database, either version 1 or 2 of the database is incorporated (Giancola et al. 2018). For general event detection, this is sufficient, however, a crucial advance of the v2 version of the Soccernet database is the expansion of manual annotations made in the dataset and further distinction between classes (Deliege et al. 2021). This refinement allows for the inclusion but also differentiation of what used to be “fouls” into 4 different classes; “foul”, “yellow card”, “red card” and “yellow -> red card”. This change allows for multiclass classification of foul events, as foul events do not only consist of just a foul, but they can consist of many more such as listed above.

However, many fail to take into account that tackles and challenges can also be legitimate, resulting in no foul. Correct tackles and challenges are performed frequently in a football match. While they may appear similar to actual fouls, it is important for the model to possess the ability to differentiate and classify these valid tackles. Failure can result in the model wrongly predicting numerous tackles that were not fouls as fouls.

The recently released Soccernet-MVFouls dataset is a specialized subsidiary dataset from the main Soccernet dataset. This new dataset focuses specifically on fouls in football. this dataset consists of 500 broadcasts from the main Soccernet database, where only the foul actions were extracted. Consequently, a total of 3,901 actions were extracted from these 500 broadcasts (Held et al. 2023).

It is worth noting that each individual action within the dataset can comprise a minimum of two clips and a maximum of four different views of the same action. This is due to the inclusion of replays from various camera angles. To ensure the accuracy and reliability of the dataset, each of the 3,901 actions was manually annotated by a professional referee. These annotations encompassed 10 distinct properties, including the type of action displayed (such as push, pull, or tackle), the classification of the offense (indicating whether an offense had been committed), and the severity of the action (ranging from no card, yellow card, to red card) (Held et al. 2023).

The inclusion of these annotations within the dataset allows for significant enhancements in the precision of foul classifications. This enables a model to recognize that not all fouls possess identical attributes and to effectively differentiate between fouls, bookable fouls, and correctly executed tackles. This distinction is particularly valuable since many previous studies, particularly those utilizing other Soccernet databases, lack a specific class for correctly performed challenges. Consequently, models trained without a designated "no foul" class may struggle to identify correctly executed tackles that do not warrant a foul, and may only be capable of annotating new data with foul labels.

Overall, accurate differentiation of bookable offenses is crucial for the model as these offenses can have significant implications on match outcomes. Therefore, the aforementioned modifications made from the parent Soccernet database to the Soccernet-MVFouls database are expected to yield substantial improvements in the automatic classification of foul events in football using deep learning techniques. The inclusion of differentiated foul event annotations in the training data sets the Soccernet-MVFouls database apart from other versions of the Soccernet database.

1. **Method**

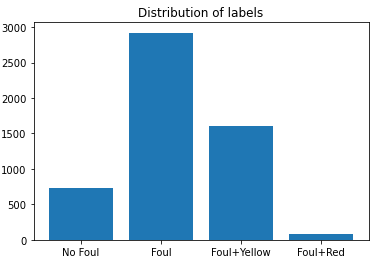
The data used in this thesis is the Soccernet-MVFoul database. This database contains 3901 actions from 500+ broadcasts of top European football, including annotations of 10 different properties, manually annotated by a professional referee (Held et al. 2023). The Soccernet-MVFoul database provides a significant advantage over the other databases as it only includes data containing foul actions and the inclusion of actions in which a challenge has been attempted but no foul was made, this addition is especially important as a model without such data would coil potentially not be able to predict that a correctly placed tackle is not a foul, it would classify any tackle or challenge as a foul (with potentially a booking).

However, it is important to acknowledge that not all data from the Soccernet-MVFouls dataset was usable due to missing annotations. Due to the lack of expertise to fill in these annotations, the decision was made to exclude these actions from the dataset. Additionally, one of the annotations, referred to as "offense," categorized actions as "Offence," "No offense," or "Between." In the case of actions annotated as "Between," indicating that the action fell somewhere between being an offense and not being an offense, it was chosen to consider them as "Offense." This decision was made to increase the size of the dataset since the Soccernet-MVFouls dataset is relatively small, containing only 3,901 actions, which would be reduced to approximately 3,000 actions after removing the unusable data (Held et al. 2023).

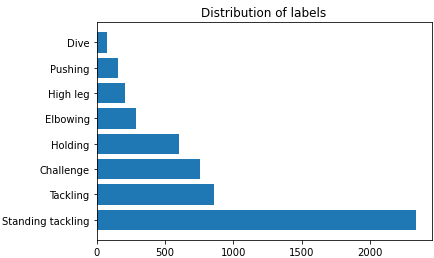
It is crucial to note that another annotation, namely "Severity," also played a role in determining whether actions annotated as "Between" were counted as offenses. The severity annotation indicated whether the action constituted a regular foul or a bookable offense. The values 1, 3, and 5 were used, with 1 representing a foul, 3 representing a yellow card offense, and 5 representing a red card offense. Similar to the offense annotation, values 2 and 4 represented actions falling in between categories. Therefore, if an action was annotated as "Between" in the offense class, it would only be considered an offense if it also had a severity annotation, as the data without the severity annotation could not be utilized.

For the severity annotation, the decision to scale down the annotations, where 1 and 2 represented a foul, 2 and 4 represented a yellow card offense, and 5 remained unchanged as a red card offense was made. These modifications were implemented to increase the amount of available data due to the relatively small nature of the Soccernet-MVFouls database mentioned earlier.

Another notable distinction between the original paper on the Soccernet-MVFouls dataset and this thesis lies in the utilization of single-view and the exclusion of the "action class" annotation. The action class annotation assigns a label to the type of action being performed, such as "high leg," "pushing," or "standing tackle." In the original paper, the model aimed to predict both the severity of the action and the action type. However, upon examining the distribution of the severity and action class annotations, it became apparent that certain combinations lacked sufficient data (See figure 1&2). In order to address this issue and ensure an adequate amount of data for analysis, this thesis will solely focus on single-view data, thereby disregarding the action class annotation. This approach allows for a more manageable dataset with sufficient data for training and analysis.



**Figure 1**: Barplot displaying the distribution of the severity labels in the dataset



**Figure 2**: Barplot displaying the distribution of the action class label in the dataset

The presence of a highly imbalanced dataset can adversely affect the training process, potentially leading to overfitting of the model on combinations with abundant data. This notion is corroborated by Hensman et al. (2015), who examined the influence of imbalanced training data on Convolutional Neural Networks (CNNs) and demonstrated that imbalanced training data significantly impairs performance compared to a more balanced dataset. Consequently, this thesis will disregard the action class annotation and exclusively utilize the severity annotation as the label for the models developed in this thesis. This approach aims to mitigate the imbalanced dataset issue and potentially enhance the obtained results.

Held et al. (2023) proposed the utilization of multiview analysis to classify foul events, incorporating multiple perspectives such as live-feed and available replays of the same action to enhance the classification process, as it suggested that the multiview approach can outperform single-view methods. However, this thesis will focus exclusively on the single-view approach for classifying foul events due to the following reasons:

Increasing dataset size: By considering different clips of each action as separate actions, more data can be added to the dataset. Since these additional clips are not required in the single-view analysis.

Investigating single-view performance: This thesis aims to explore the results obtained through the single-view approach in the current experimental setup for classifying foul events. The modifications made to the methodology of this thesis may potentially improve results by reducing data imbalance and simplifying the model's complexity.

Additionally, it is important to note that there is no overlap in actions between the train, validation, and test sets.

In order to distinguish between valid tackles and fouls, as well as to determine fouls and bookable fouls, this thesis will employ multiclass labeling with four distinct labels. The labels will be represented in the form of [0,0,0,0], where each label corresponds to one of the following categories: "No foul," "Foul," "Foul+Yellow," and "Foul+Red."

This thesis aims to explore the feasibility of automatically classifying foul events in football by comparing three different models. The models selected for comparison are r3d\_18, mc3\_18, and r2plus1d\_18, which are readily available from the PyTorch library. The choice of utilizing resnet models is motivated by their exceptional performance in video classification tasks and their frequent application in other football classification studies, such as those conducted by Rongved et al. (2020) and Tomei et al. (2021). Furthermore, in order to compare the results for the first sub-research question, two of the models, namely r3d\_18 and 2plus1d\_18, were chosen. The mc3\_18 model was specifically selected to facilitate a direct comparison with the findings of Held et al. (2023) and address the second sub-research question effectively.

In this thesis, the selected models will be utilized in conjunction with transfer learning. The pre-trained weights of the models, which were trained on the widely employed KINETICS V400 dataset, will be obtained from PyTorch. By employing transfer learning, the model training process can be accelerated compared to training from scratch, and there is a potential for improved results (Sen et al. 2022).

The first sub-research question aims to assess the impact of data augmentation on a model's performance, taking into consideration the aforementioned limitations of the dataset and the concept of data bias associated with data augmentation. Xu et al. (2020) conducted a study exploring methods to enhance the utilization of data augmentation. It was noted that the main limitation of data augmentation is data bias, referring to the augmented data being quite different from the original one, leading to a suboptimal performance (Xu et al. 2020). In order to address this sub-research question, the performances of all the models will be compared both with and without the application of data augmentation.

To effectively achieve the primary objective of data augmentation, which is to expand the dataset, it is crucial to employ augmentations that do not render the data unusable. Consequently, this thesis will incorporate several augmentations to enhance the dataset, including horizontal flip, resized crop, color jitter, and random rotation. Horizontal flip serves as a valuable augmentation because it flips the image horizontally, accommodating situations where tackles can come from either side. This can potentially reduce overfitting associated with tackles originating from a specific direction. Random rotation introduces a random rotation of the images by a specified degree. This augmentation proves useful when employing a low degree value, as it ensures the model's resilience to minor rotational variations from sources such as the cameras. Resized crop standardizes the size of images in the dataset while preserving important visual features. By reducing noise within the data, this augmentation contributes to improving overall data quality. Lastly, color jitter is employed to randomly alter the brightness, contrast, saturation, and hue of the images. This augmentation is particularly significant as it ensures that the model remains unaffected by variations in the color of the player's jerseys. By reinforcing the notion that color should not impact the model's performance, this augmentation helps maintain its robustness. This analysis will provide insights into the extent to which data augmentation affects the models' performance, considering the challenges posed by data bias.

In line with the common practices in football video classification research, this thesis will employ the Adam model as the optimizer and the cross-entropy loss function. The Adam optimizer has demonstrated its effectiveness in training deep learning models, including those applied to football video classification tasks. Similarly, the cross-entropy loss function is widely used for multi-class classification problems and has been extensively applied in previous research on football video classification. By utilizing these established components, this thesis aims to align with the existing methodologies and ensure consistency with prior studies in the field (Rana et al. 2023; Giancola et al. 2021; Rongved et al. 2020; Tomei et al. 2021).

To address the primary limitation arising from the imbalanced nature of the dataset, L2 regularization will be employed in this thesis. Kamalov et al. (2020) demonstrated the efficacy of L2 regularization in mitigating the challenges associated with imbalanced data in deep learning. By incorporating L2 regularization, the weights of the model can be appropriately decayed, which helps to prevent overfitting and improves the generalization ability of the model (Kamalov et al. 2020).

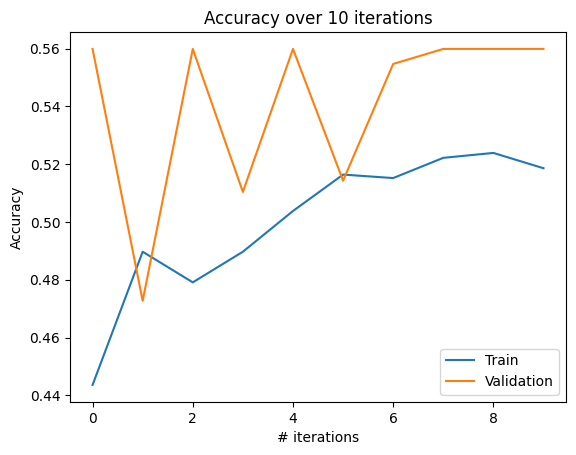
In order to further address the issue of class imbalances, class weights will be employed as an additional remedy in this thesis. The class weights will be determined based on the distribution of classes in the dataset (See figure 1) and will be normalized to suit the loss function. By assigning different weights to each class, based on their distribution, the model can give more importance to minority classes, thereby enabling better prediction performance (Asundi et al. 2022; Phan et al. 2020). The use of class weights in handling imbalanced datasets is supported by previous studies such as Asundi et al. (2022) and Phan et al. (2020). These studies highlight the effectiveness of assigning higher weights to minority classes to address the challenges posed by imbalanced data.

For model evaluation, the models will undergo a training process consisting of 10 epochs, with this maximum number of epochs limited by memory constraints. The evaluation of the models involves assessing their accuracy and loss on both the validation and training sets, aiming to identify in which epoch the model exhibits the lowest accuracy and loss values on both sets. Subsequently, the selected models were subjected to a performance comparison using the F1 score and weighted F1 score on the test set. The inclusion of both the F1 score and the weighted F1 score in the analysis aimed to assess the model's performance based on the F1 score while taking into consideration the class distribution by utilizing the weighted F1 score.

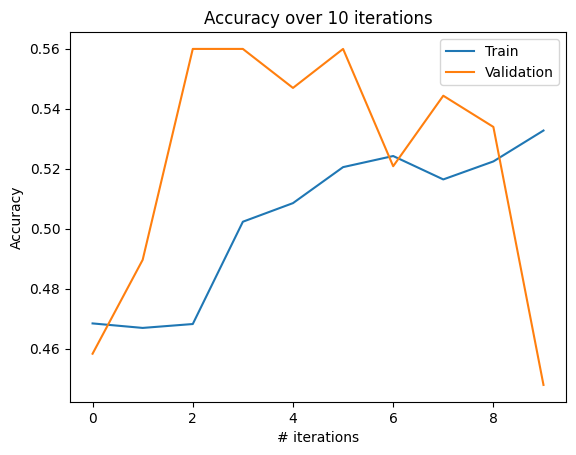
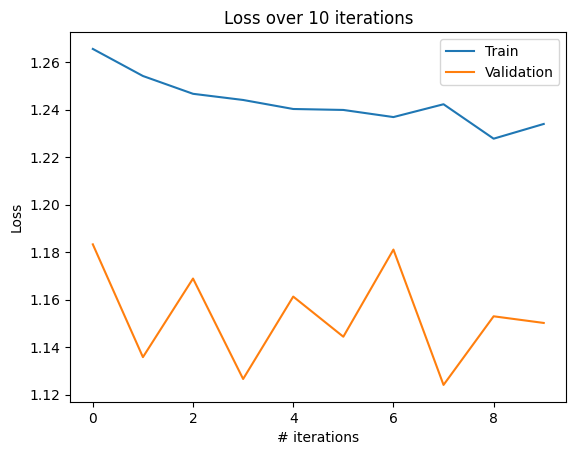
1. **Results**

The performances of all the models, considering both the scenarios with and without data augmentation, will be presented in Figures 3-8. These figures will display the accuracy and loss metrics for each model, allowing for a comparison of performance between the two conditions.

By examining the accuracy and loss plots for each model, the iteration that exhibits the best performance in terms of both accuracy and loss will be identified. This assessment will be conducted for both the training and validation sets, enabling a comprehensive evaluation of the models' performance under different conditions. After identifying the best iteration of each model under both conditions (with and without data augmentation) based on the accuracy and loss metrics, these selected models will be further evaluated using the test set. The performance of each model will be assessed on the test set, allowing for a comparison of their performance in a real-world scenario. By comparing the performance of the models under both conditions (with and without data augmentation) on the test set, the thesis can determine the extent to which data augmentation influences the model's performance. Additionally, evaluating the performance of the mc3\_18 model in relation to the other models allows for an assessment of its effectiveness and provides insights into its comparative performance.

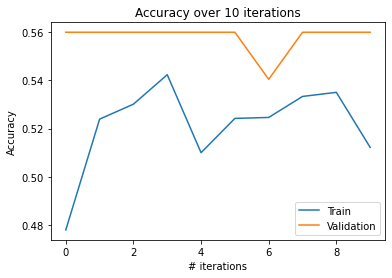
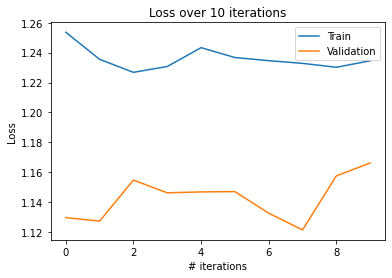


**Figure 3**: Accuracy and loss graphs of r3d\_18 model trained for 10 epochs with data augmentation

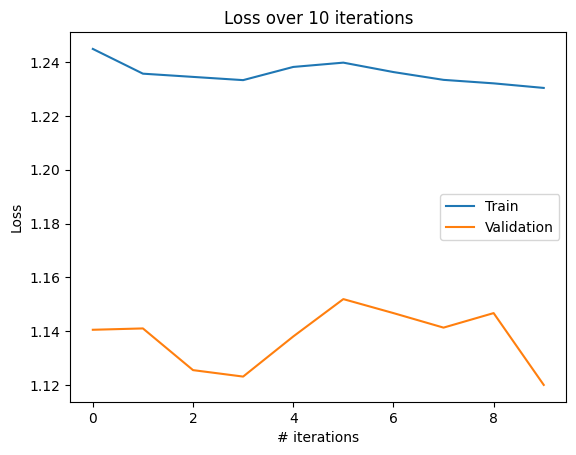
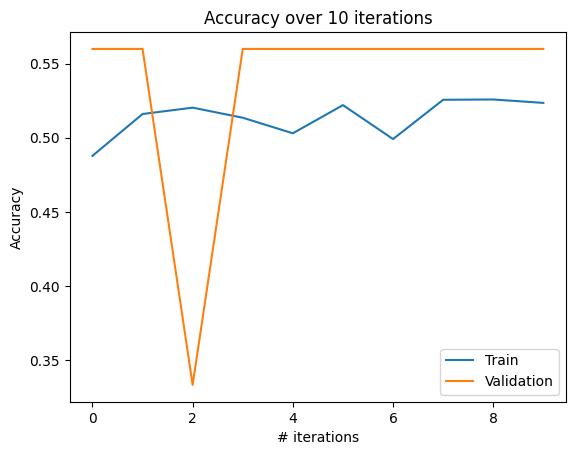


**Figure 4**: Accuracy and loss graphs of r3d\_18 model trained for 10 epochs without data augmentation

The analysis of the r3d\_18 model's results reveals several observations. Firstly, there is a relatively small difference in loss between the two conditions (with and without data augmentation). However, the results without data augmentation appear more scattered, with larger dips and rises. The decline in accuracy observed on the validation set after the 7th iteration may suggest the occurrence of overfitting to the training data. On the other hand, the results with data augmentation demonstrate a more consistent upward trend. Based on these findings, the last iteration appears to offer the best combination of accuracy and loss for the r3d\_18 model with data augmentation. However, for the model without data augmentation, the 6th or 5th iteration seems to indicate the best performance. To ensure a thorough exploration and selection of the best model, both iterations will be further investigated, and the model that yields the optimal performance will be reported.

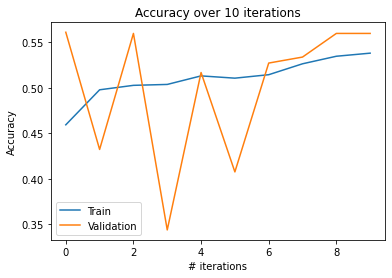
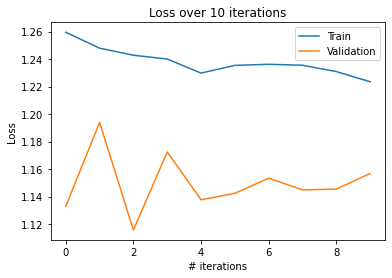


**Figure 5**: Accuracy and loss graphs of mc3\_18 model trained for 10 epochs with data augmentation

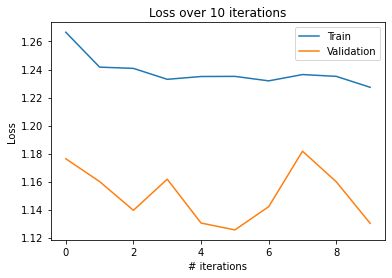
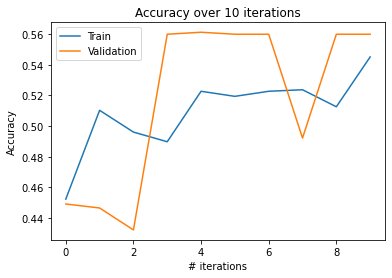


**Figure 6**: Accuracy and loss graphs of mc3\_18 model trained for 10 epochs without data augmentation

The evaluation of the mc3\_18 model reveals notable patterns in its performance under both conditions. In both scenarios, the accuracy of the validation set appears to stagnate at a value of 0.5599 across multiple iterations, with only a temporary dip observed for a single iteration. This suggests that the model may be trapped in a local minimum and is unable to escape from it. Comparing the two conditions, the model without data augmentation exhibits a slightly more desirable trend, with an upward trajectory in accuracy and a downward trajectory in loss, demonstrating fewer deviations compared to the model with data augmentation. The model without data augmentation appears to yield the best results at the last iteration. However, to ensure that overfitting is not occurring, the 3rd iteration will also be tested as it displays similar performance trends. Similarly, for the model with data augmentation, both the 3rd and 7th iterations will be tested to identify the best-performing model.



**Figure 7**: Accuracy and loss graphs of 2plus1d\_18 model trained for 10 epochs with data augmentation



**Figure 8**: Accuracy and loss graphs of 2plus1d\_18 model trained for 10 epochs without data augmentation

The evaluation of the 2plus1d\_18 model yields comparable results to the previous model. In both conditions, the observed trends exhibit desirable characteristics but also include deviations and appear to reach a ceiling. Notably, the validation set achieves a slightly higher ceiling in performance, albeit by a small margin. This minor difference in performance between the training and validation sets could potentially be attributed to slight variations in their respective distributions. The evaluation of the model with data augmentation will include testing the final two iterations to assess their performance on the test set. Since the results of these iterations are not significantly different, it is necessary to determine which iteration performs better. Similarly, for the model without data augmentation, the 5th and last iteration will be tested. Although the 5th iteration exhibits a lower accuracy on the training set compared to the last iteration, it still demonstrates promising results across other metrics.

| Data augmentation Models | Weighted *F*1 | *F*1 |
| --- | --- | --- |
| r3d\_18 | 0.45 | 0.570 |
| Yes mc3\_18 | 0.47 | 0.614 |
| r2plus1d\_18 | 0.46 | 0.574 |
| r3d\_18 | 0.48 | 0.562 |
| No mc3\_18 | 0.47 | 0.614 |
| r2plus1d\_18 | 0.47 | 0.616 |

**Table 1**: Weighted *F*1 scores and Average *F*1 scores (rounded up) of best-performing models classifying foul events on the test set. (top) models trained with data augmentation, (bottom) models trained without data augmentation.

Table 1 presents the results of all the models under both conditions, showcasing the F1 scores and weighted F1 scores. The weighted accuracy is provided to account for the unbalanced distribution of classes within the dataset. The F1 scores, which provide a measure of the model's precision and recall, demonstrate minimal differences among all the models. In general, the F1 scores are similar across the models, indicating comparable performance in terms of overall accuracy and effectiveness in classifying foul events. Notably, the only substantial difference observed is between the two r2plus1d\_18 models. In this case, the model without data augmentation exhibits a slightly higher accuracy by a margin of 0.04 compared to the model with data augmentation. Similarly, the weighted F1 scores demonstrate a high degree of similarity across the different conditions. Notably, the condition without data augmentation tends to yield slightly higher weighted F1 scores compared to the condition with data augmentation.

1. **Discussion**

This thesis has examined the automatic classification of foul events in football using deep learning, alongside two sub-research questions which explored to what degree does data augmentation affect a model’s performance and to what extent does the introduction of a new model increase results. The results obtained in this thesis are highly promising. Despite considering the imbalanced distribution of classes, the weighted F1 scores achieved by the models are satisfactory and surpass previous research efforts in this domain. These findings underscore the effectiveness of the proposed approach in accurately classifying foul events in football.

The results to the first sub-research question indicate only marginal improvements without employing data augmentation. This suggests that the application of data augmentation techniques did not significantly enhance the model's performance in the specific task of foul event classification. One potential explanation for this observation is that data augmentation is primarily employed to introduce diversity into the training data, thereby improving the model's ability to generalize and perform well on unseen data. However, in the context of football foul events, the visual characteristics of these events are often quite similar. Fouls share common visual attributes, differing primarily in the specific action taken to commit the foul (e.g., push, pull, tackle). As a result, augmenting the data to introduce significant variations may lead to data bias, where the augmented instances become too dissimilar from the original data, ultimately hampering the model's performance. The lack of distinct visual variations within the foul event classes may hinder the efficacy of augmentation techniques, potentially resulting in a counterproductive approach.

The second sub-research question aimed to assess the impact of introducing a new model on the performance of foul event classification. The results obtained in this thesis indicate satisfactory outcomes, particularly with the mc3\_18 model achieving a performance of 47%. These findings suggest that alternative models can also be effective in the classification of foul events in football.The success of the mc3\_18 model implies the potential benefits of exploring additional models in future research endeavors. By considering newer models or models specifically tailored to the task at hand, it is plausible to expect further improvements in classification results. However, it is worth noting that the relatively high performance of the mc3\_18 model is not entirely surprising, as it shares similarities with the other models explored in this thesis.

The findings of this thesis indicate superior performance compared to the results reported by Held et al. (2023). Specifically, in relation to the two models shared between the two studies, Held et al. achieved a weighted F1 score of 0.28 for the r3d\_18 model and 0.34 for the r2plus1d\_18 model. In contrast, this thesis achieved notably higher weighted accuracy scores, with a difference of 0.20 and 0.13 respectively. Furthermore, the mc3\_18 model, which was not utilized in Held et al.'s research, demonstrates a similarly high performance in this thesis of 0.47.

The research conducted by Spitz et al. (2018) on the impact of video speed on the decision-making process of sports officials provides valuable insights into the accuracy levels achieved by referees in assessing foul events. The reported accuracy rates of approximately 61% for real-time video and 63% for slow-motion video highlight the inherent challenges and limitations faced by human officials in accurately identifying and classifying fouls. Considering that the weighted accuracy of the deep learning model in this thesis reaches around 47%, it is important to contextualize these results. The promising results obtained in this thesis, with a weighted accuracy approaching the accuracy levels of human referees, indicate that the automatic classification of foul events using a deep learning approach holds significant potential.

Furthermore, it is important to assess the specific errors made by the model in order to gain a comprehensive understanding of its performance. Notably, wrongly classifying a no-foul as a foul may have limited consequences, compared to misclassifying a foul as a red card. Hence, exploring indicators that quantify the proximity of the model's predictions to the true action, may offer valuable insights into the model's potential for improvement.

The findings of this thesis suggest that the model may have been prone to overfitting, particularly towards the majority classes. This assertion is supported by the notable disparity observed between the F1 score and weighted F1 score, implying that certain classes either received no predictions or were inaccurately predicted. Moreover, the presence of ceilings in figures 3-8, with all reaching a maximum of approximately 0.56, indicates the existence of minima that the model struggled to surpass.

It is crucial to acknowledge the limitations that could have influenced the outcomes of this thesis. One notable limitation pertains to the shift from a multi-view to a single-view approach. Although the intention behind this change was to increase the available data, it could’ve worsened the issue of class imbalance within the dataset. By duplicating the number of data instances, the majority classes received more substantial benefits compared to the minority classes. It is important to note, however, that this imbalance alone cannot solely account for the suboptimal results obtained. Numerous measures were implemented to address the class imbalance, including data augmentation, regularization techniques, and incorporating class distribution in the loss function. Therefore, while the imbalance introduced by the transition from multi-view to single-view may have had some impact, other factors should be considered in understanding the overall performance of the models.

Another important limitation to consider is the effect of data augmentation on the class imbalance within the dataset. Although data augmentation is intended to enhance the model's performance by generating additional diverse samples, it could’ve worsened the issue of class imbalance. The random selection and augmentation of data tend to produce more samples of the majority classes, while the minority classes may receive less representation. Consequently, the augmented dataset becomes even more unbalanced, potentially impacting the model's ability to accurately classify the minority classes. While data augmentation is beneficial in preventing overfitting, the inherent class imbalance of the dataset may have influenced the efficacy of this technique.

Furthermore, it is important to acknowledge that the primary objective of this thesis was not to replicate the results of previous studies or solely introduce a new condition for comparison. Rather, the aim was to explore the automatic classification of foul events in football using a deep-learning approach and examine the influence of specific conditions on the results. Therefore, direct comparisons with the results of original papers are not completely feasible without following the same experimental procedures and introducing only one new condition at a time to assess its impact on the outcomes.

Not all limitations could be anticipated and addressed within the scope and capabilities of this thesis. Despite efforts to mitigate potential limitations, there may be certain factors or constraints that were beyond the scope of this thesis. It is important for future research to consider and address these limitations to further enhance the understanding and applicability of the proposed approach.

Future research can focus on improving the results through various approaches. One potential avenue for improvement is enhancing the quality of the dataset. As with any dataset, higher-quality data can facilitate better learning outcomes for the model. Some of the older video clips in the dataset may suffer from graininess or pixelation, which can hinder the model's ability to learn effectively. Exploring techniques such as applying smoothing filters or other image enhancement methods could be beneficial in improving the quality of the videos. However, due to technological constraints, this approach was beyond the scope of the current bachelor thesis.

Another approach to consider is increasing the amount of data available for the minority classes. Given the significant class imbalance in the dataset, augmenting the data specifically for the underrepresented classes can help address this issue.

A potential approach to improve the results is the addition of oversampling techniques. As mentioned earlier, the transition from multi-view to single-view increased the data volume but also exacerbated the existing class imbalance. By applying oversampling specifically to the minority classes, it helps to mitigate the imbalance by increasing the representation of these classes in the dataset. While methods such as data augmentation, regularization, weight decay, and class weighting have been implemented to address the class imbalance, research suggests that oversampling can be an effective strategy to further improve the performance of models trained on imbalanced datasets. By providing additional training examples for the minority classes, oversampling can help the model learn more robust representations and potentially achieve better results.

1. **Conclusion**

In conclusion the results obtained in this thesis, while not reaching the desired level for practical use, show promise for further exploration in this area. Furthermore, the investigation into the impact of data augmentation on model performance has revealed that its effectiveness can vary depending on the task. In this specific context, data augmentation did not yield significant improvements and may even introduce data bias. This highlights the importance of carefully considering the applicability of data augmentation techniques in different domains. Moreover, the exploration of introducing new models to the classification task has demonstrated that previously unexplored models can deliver comparable results to the models examined in this thesis. This suggests that further advancements in the field of automatic classification of foul events in football can be achieved by leveraging newer and potentially more suitable models.

Overall, this thesis serves as a contribution to future research endeavors. By addressing the limitations identified, it can strive toward a more accurate and reliable automatic classification of foul events in football.

**References**

1. Published by Statista Research Department, & 5, O. (2022, October 5). Revenue of the big five by country Europe 2021. Statista. Retrieved February 27, 2023, from https://www.statista.com/statistics/261218/big-five-european-soccer-leagues-revenue/
2. Mustapha, I. (2023, February 17). Lee Mason leaves role as Premier League referee following var controversy in arsenal v brentford. Eurosport UK. Retrieved February 27, 2023, from https://www.eurosport.co.uk/football/lee-mason-leaves-role-as-premier-league-referee-following-var-controversy-in-arsenal-v-brentford\_sto9401041/story.shtml
3. Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., & Van Droogenbroeck, M. (2023). VARS: Video Assistant Referee System for Automated Soccer Decision Making from Multiple Views. arXiv preprint arXiv:2304.04617.
4. Vidal-Codina, F., Evans, N., El Fakir, B., & Billingham, J. (2022). Automatic event detection in football using tracking data. Sports Engineering, 25(1), 18.
5. Sen, A., & Deb, K. (2022). Categorization of actions in soccer videos using a combination of transfer learning and Gated Recurrent Unit. ICT Express, 8(1), 65-71.
6. Sangram Singh Rana, A. (2023). Event Detection in Football using Graph Convolutional Networks. arXiv e-prints, arXiv-2301.
7. Giancola, S., & Ghanem, B. (2021). Temporally-aware feature pooling for action spotting in soccer broadcasts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4490-4499).
8. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
9. Thamaraimanalan, T., Naveena, D., Ramya, M., & Madhubala, M. (2020). Prediction and classification of fouls in soccer game using deep learning. Ir. Interdiscip. J. Sci. Res, 4, 66-78.
10. Silvio Giancola, Mohieddine Amine, Tarek Dghaily, Bernard Ghanem (2018). SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. https://arxiv.org/abs/1804.04527.
11. Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, Marc Van Droogenbroeck (2021). SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. https://openaccess.thecvf.com/content/CVPR2021W/CVSports/papers/Deliege\_SoccerNet-v2\_A\_Dataset\_and\_Benchmarks\_for\_Holistic\_Understanding\_of\_Broadcast\_CVPRW\_2021\_paper.pdf
12. Hensman, P., & Masko, D. (2015). The impact of imbalanced training data for convolutional neural networks. Degree Project in Computer Science, KTH Royal Institute of Technology.
13. Rongved, O. A. N., Hicks, S. A., Thambawita, V., Stensland, H. K., Zouganeli, E., Johansen, D., ... & Halvorsen, P. (2020, December). Real-time detection of events in soccer videos using 3D convolutional neural networks. In 2020 IEEE International Symposium on Multimedia (ISM) (pp. 135-144). IEEE.
14. Tomei, M., Baraldi, L., Calderara, S., Bronzin, S., & Cucchiara, R. (2021, January). Rms-net: Regression and masking for soccer event spotting. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 7699-7706). IEEE.
15. Xu, Y., Noy, A., Lin, M., Qian, Q., Li, H., & Jin, R. (2020). Wemix: How to better utilize data augmentation. arXiv preprint arXiv:2010.01267.
16. Kamalov, F., & Leung, H. H. (2020, November). Deep learning regularization in imbalanced data. In 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI) (pp. 1-5). IEEE.
17. Asundi, R. V., Prakash, R., & Kumar, K. Class Weight technique for Handling Class Imbalance.
18. Phan, T. H., & Yamamoto, K. (2020). Resolving class imbalance in object detection with weighted cross entropy losses. arXiv preprint arXiv:2006.01413.
19. Spitz, J., Moors, P., Wagemans, J., & Helsen, W. F. (2018). The impact of video speed on the decision-making process of sports officials. Cognitive Research: Principles and Implications, 3(1), 1-10.