Methodological Challenges in Rare Event Prediction: A Case Study of Time Series Modeling for Sports Injuries

By

Momchil Bachvarov

Submitted to Central European University – Private University Undergraduate Studies

In partial fulfilment to the requirement for the Bachelor of Science in Data Science and Society

Supervisor: Gábor Békés

Vienna, Austria
2024

Copyright Notice

For bibliographic and reference purposes this thesis/dissertation should be referred to as: Bachvarov, M. 2025. Methodological Challenges in Rare Event Prediction: A Case Study of Time Series Modeling for Sports Injuries. BA thesis, Undergraduate Studies, Central European University, Vienna.

Abstract

Injuries are an inevitable part of competitive sports but thanks to modern machine learning approaches and the vast amounts of data available, new opportunities to analyze and potentially predict injury risk have emerged. This thesis investigates the methodological challenges of predicting rare events in time series data through a case study on sport injuries. A dataset collected from a track and field team over a seven-year period, was used to train several machine learning methods – including LSTM, GRU and hybrid CNN+GRU architectures. Rather than focusing solely on performance optimization the study explored the methodological and theoretical challeges of predicting rare events. Each modeling decision such as class imbalance handling, feature representation, and validation strategy is analyzed in term of its impact. The findings highlight the fragility of rare event prediction pipelines and emphasize the importance of methodological choices over the model's complexity. The goal of the work is to contribute to a better understanding of practical and theoretical limits of machine learning for injjury prediction, while exploring the challenges of rare event prediction in a time series context.

Authors Declaration

I, the undersigned, **Momchil Bachvarov**, candidate for the BA degree in Data Science and Society declare herewith that the present thesis titled “Methodological Challenges in Rare Event Prediction: A Case Study of Time Series Modeling for Sports Injuries” is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person’s or institution’s copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of education for an academic degree.

Vienna, 26 May 2025

Momchil Bachvarov

Table of Contents

List of Figures

Introduction.

Rare event prediction is one of the most challenging tasks in applied machine learning. Utilized in many industries such as finance, healthcare, sports, and manufacturing, being able to forecast a low-frequency but very impactful event, such as a default, failure, or injury, offers substantial value. However, the nature of these events makes them hard to detect and predict. They commonly pose challenges such as highly imbalanced data, noisy or weak signals, and patterns that may change over time. All these factors create methodological challenges that the standard machine learning techniques are not designed to tackle.

In the context of time series, these challenges are even further amplified. Rare events not only happen infrequently but also occur in sequences of temporally dependent data. This introduces additional tasks such as preserving the time dependencies in the data, detecting subtle changes over time, and avoiding information leakage in the data from the future to the past. Many data preprocessing and modeling strategies have been proposed - including oversampling techniques, specialized loss functions, class weights, and hybrid model architectures, but the challenge still remains difficult to evaluate, generalize, and reproduce.

This thesis looks at these methodological challenges through the lens of sports injury prediction - a case where rare events (injuries) occur within continuous data of athlete observation. While the case is specific the paper reflects the broader limitations and unresolved questions in rare events prediction for time series.

While machine learning has become very effective at large-scale prediction tasks, it continues to struggle when applied to rare event prediction tasks - especially with time series data. These difficulties are not only technical but also methodological and conceptual. What makes a good model in rare contexts? How should time dependencies be preserved during training and evaluation? What are the implications of imbalance, thresholding, and sampling techniques on generalizability and interpretability? These questions remain unclear despite the growing body of work.

Studies on rare event prediction exist in different domains. Examples of Injury prediction (Lövdal et al., 2021), fraud detection (Dal Pozzolo et al., 2015), medical diagnostics (Johnson et al., 2016), and equipment failure forecasting (Zhao et al., 2019) consistently point to core challenges: severe class imbalance, noisy or sparse signals, unstable decision thresholds, and difficulty generalizing across different contexts. To solve such issues researchers have proposed a range of strategies including oversampling methods like SMOTE (Chawla et al., 2002), cost-sensitive learning (Elkan, 2001), temporal ensembling (Laine & Aila, 2017), and hybrid architectures that combine statistical and deep learning features (Zhang et al., 2021).

In the context of time series the temporal dependencies of the data add more complexity. Approaches such as Recurrent Neural Networks (RNNs), convolutional neural networks, and attention-based models have been applied to sequential data (Hochreiter & Schmidhuber, 1997; Bai et al., 2018). These methods have proven to be effective but results often decrease when rare events are sparsely distributed with limited positive samples.

In the field of sports injuries, Løvdal et al. (2021) conducted one of the most comprehensive studies. They apply a machine learning model to data from track athletes achieving high predictive performance in forecasting injuries. Their combination of raw time series inputs in combination with engineered features helps them achieve these results. However, even with such comprehensive modeling pipelines, reproducing similar results can prove to be difficult in practice. Small differences in implication, data, preprocessing, or sampling strategies can lead to significant variations in outcome. This raises broader questions about the robustness and generalizability of rare event prediction models in such specific cases.

This thesis is closely related to the work of Løvdal et al. (2021) as an example of the challenges in predicting rare events in time series data based on building a model on the same data used by the researchers. This thesis narrows down on this broader research problem: What are the theoretical and methodological challenges when trying to predict rare events in time series data? By focusing on the specific question, the aim is not to optimize the performance on a specific dataset but rather to critically examine the assumptions, design choices, and limitations that come with rare event modeling. To illustrate the discussion concretely, the thesis shows an applied example of injury prediction using athlete monitoring data. The example demonstrates many of the general challenges in rare event time series modeling.

To explore these challenges, this thesis goes through building a pipeline for injury prediction based on the dataset. Each step in the pipeline - from data preprocessing and feature engineering to model selection, class imbalance, and evaluation strategy- is examined in detail. Techniques such as class weighting, focal loss, threshold tuning, and hybrid architectures are applied and

compared, not only in performance but also in methodological implications. Rather than focusing on developing a single high-performing model, different models were tested and evaluated showing the trade-offs they bring. Through this process, the thesis aims to highlight which methods are robust and how modeling outcomes can vary depending on the choices you make in each step.

This thesis starts with a literature review that presents relevant literature on event prediction and time series modeling with a focus on methodological challenges. The next, Methodology chapter, outlines the data, preprocessing techniques, modeling strategies, and evaluation methods, used. This is followed by the results chapter that presents the findings from the modeling experiments, organized mode by model. The subsequent chapter is a reflection of the modeling process and the results achieved. The final chapter concludes the thesis by summarizing the main contributions and proposing a direction for future work.

1 Literature Review

This literature review goes over the theoretical and methodological challenges relevant to this thesis and the papers that support them. While the applied context is injury prediction in sports, the broader aim is to explore common modeling techniques used in rare event prediction. To begin the review the study by Lövdal et al.(2021), which provides both the dataset and an important benchmark, is discussed. This is followed by a review of general supporting literature on the topic of rare event prediction in time series contexts.

Lövdal et al.'s (2021) study on injury prediction with machine learning is the paper from which I have taken my data. Their study is one of the most comprehensive in the field of injury prediction with machine learning. In their study, the authors have developed an XGBoost classifier on a dataset of track athletes. They introduce two data preprocessing approaches: a day approach where the seven days preceding an event(healthy or injured) were described by lagged features, and a week approach where training load features were aggregated for the three weeks prior to the event.Another preprocessing approach was the athlete-specific z-score normalization they implemented. The prediction accuracy results they got from the study were significant, especially in the field of sports science.

However, several limitations of the Lövdal et al. (2021) study become apparent upon closer examination. Even though the study incorporates lagged features when using the day approach, the model they use - XGBoost is a tree-based, non-sequential model that restricts the algorithm's ability to learn from cumulative patterns over time. The model processes the lagged features as

static inputs without a mechanism to model the sequential dynamics or memory of the prior states beyond the 7-day window, which limits its capacity to detect workload accumulations and does not take advantage of the large dataset, tracking the athletes over a 7-year period.

Secondly, while the study acknowledges the rare occurrence of injuries and addresses the class imbalance in the dataset, the bagging approach they apply remains limited. The study does not explore alternative balancing mitigation methods such as cost-sensitive learning or anomaly detection frameworks which have proven to be effective (Shaylika et al, 2024). Additionally, their evaluation consists mainly of aggregate classification metrics. An approach like an in-depth analysis of the model's behavior in the injury class looking at false negatives and false positives could have provided critical insights.

The task of predicting sports injuries shares similar challenges as other domains like fraud detection, industrial equipment failure, or rare disease detection (Shaylika et al, 2024). In all of these cases, the target event is very rare but has a huge impact. It is commonly caused by an accumulation of factors over time, which makes the task of modeling it crucial but very demanding at the same time.

Shaylika et al, (2024) provide a comprehensive study on rare event prediction. They outline the main challenges and approaches when dealing with rare events in time series data. Among the most significant issues identified are class imbalances, where most of the observations belong in a negative (non-event) class. A case similar to the track athletes data where over the period of 7 years less than 5% of the cases an athlete is injured. This imbalance often leads to machine

learning algorithms favoring the majority class and presenting deceiving accuracy results if not handled correctly. The authors emphasize that metrics such as accuracy are misleading in similar cases and metrics like precision, recall, F1 score, and PR-AUC should be prioritized.

Another challenge Shaylike et al. (2024) highlighted is temporal dependency and autocorrelation. In time series data events are often not independent and the sequence of events leading up to a point can hold important predictive signals. This is why models like the XGBoost used in Lövdal et al. 's (2021) paper, which ignores these dependencies could be missing these critical warning patterns.

The challenge of rare event prediction and time series modeling has been a topic that researchers have been developing and testing for a while. Traditional practices of handling imbalanced datasets like oversampling the minority class, undersampling the majority class, or applying sampling techniques like SMOTE, have been widely used (He & Garcia, 2009). However, when it comes to time series where preservation of temporal dependencies is crucial such techniques fall short. Techniques like SMOTE, when applied to temporal data, may disrupt the sequential structure of the data and disrupt important patterns used to predict rare events.

An alternative strategy that is being explored by researchers is cost-sensitive learning, where the model is penalized more heavily when misclassifying minority examples (Krawczyk, 2016). Recent advancements in deep learning, moreparticularly in Recurrent Neural Networks (RNN) such as Long Short-Term Memory offer promising results for modeling rare events in time series data. Ramachand (2020) for example explored an ensemble LSTM approach for the

detection of rare events. He demonstrated that combining multiple LSTMs trained on different splits of the data can help with overfitting and improve minority class detection.

Recent work also shows that hybrid models - architectures that combine learned representations with engineered features bring promising results (Zhang et al. 2021). These models can help compensate for the limited signals in the data and the sparse labels, common in rare event prediction contexts by leveraging deep learning techniques with easily interpretable statistical summaries.

In conclusion, the combination of machine learning and time series modeling for rare event prediction has evolved in other fields, even in close ones like healthcare, but it remains underexplored in the field of injury prediction in sports. My thesis aims to explore the research gap and present challenges that have been faced in similar fields and how they could be handled in the field of predicting sports injuries.

2 Methodology

The following chapter provides the theoretical background and the logic behind the decisions taken during the modeling process in this thesis. The primary goal is to explore the challenges involved in predicting rare events in time series data. Each stage of the modeling pipeline is examined both in its practical application and theoretical implications.

Rather than focusing only on the successful configurations the chapter includes the range of possibilities presented to me when examining the existing literature, before taking the final decision. This includes various data preprocessing techniques, data splitting challenges, model choices, and different evaluation metrics. By presenting these steps in chronological order my aim is to make the challenges of each step more clear and the solutions more replicable.

2.1 Data

The dataset used in this thesis originates from the study conducted by Lövdal et al. (2021), which collected training and injury data from 74 high-level track team athletes competing in medium to long distances over 7 years. Each row in the data represents a single training day for a specific athlete. The columns in the datasets are different measures of the athletes training such as total kilometers run, strength training, kilometers sprinting, number of sessions, perceived recovery, and others. The dataset also includes information about the injuries of the athletes, dates, and an ID. The raw version of the dataset included 42,766 rows each of which is the data of a specific athlete for a specific day and 73 columns which are different features measured. Taking the data from an existing study made the task easier as the data was already cleaned and well structured.

2.2 Data Preprocessing

One of the initial challenges in preparing time series data for rare event prediction is designing input windows to retain temporal dependencies while avoiding redundant data and information leakage. In this thesis, a sliding window approach is used where each window consists of 20 days. The reason I have opted for 20-day windows is because too short windows may miss evolving trends in the data and too long windows increase the risk of diluting important signals in the data, especially considering the rare occurrence of injuries in the data. (Carey et al., 2018; Rossi et al., 2018). The original dataset included engineered lagged features for seven days displaying the data for the 7 days prior to a specific event and basically creating a 7 day window for each row of information in the data. I have decided to exclude these features from the data as my idea was always to train time-aware architecture in which temporal dependencies are learned directly from the sequences and such features could be unnecessary or potentially harmful.

An additional design decision that has to be made is one for the prediction target. The question is whether a period is assigned with a positive injury label if the injury occurs during the period or on the day after the window period. The first approach risks information leakage as it allows the model to "see" data from days right before the injury. On the other hand, the strict forecasting approach, where we label only if the injury occurs on the next day, may miss signals that happen a few days earlier. I have opted for a risk window approach: each 20-day window was labeled as positive if an injury occurred during a short window right after the 20-day window. This short window (between 3-5 days) approach balances the strict theory with the nature of injuries. It maintains the temporal causality by excluding the injury from the input window but also accounts that predicting an injury for a specific day might be too extreme and signaling a high-

risk period would be a more reasonable approach. The size of this window is a hyperparameter, whose influence on the results I will cover in the following results chapter.

2.3 Normalization

In contrast to many prior studies on injury prediction, this thesis did not apply any normalization or scaling to the input features during the preprocessing phase. Initial experiments explored athlete-wise normalization- where features are scaled to zero mean and unit variance for each individual athlete, similar to what Lövdal et al. (2021) adopted in their study. However, this was excluded in the final pipeline version after observing that the raw features and their absolute magnitude and patterns were clearer to the models and offered better performance. Additionally, skipping this step simplified the preprocessing pipeline.

Alternative approaches like global standardization and min-max scaling were considered but not adopted as global normalization removes the athlete-specific dynamics and compresses their inter-individual variability, while min-max scaling is more appropriate for bounded metrics or outlier-heavy features. Although normalization is a common step in time series modeling in the case of this dataset opting to not include it did not degrade the performance but the exact opposite. This choice reinforces a broader theme in the thesis - that standard methodological choices should be empirically validated and tested rather than just assumed and adopted, particularly in domains where the context and inter-individual variability may carry useful information.

2.4 Data Splitting

Creating a proper data-splitting strategy is a central methodological challenge when dealing with rare event prediction, especially when working with time series data. In cases, similar to injury prediction, standard approaches like random sampling or stratified sampling create a risk of temporal leakage, where information from the future can affect the model during training, leading to artificially inflated performance (Kelleher et al. 2015). In the dataset I am using, where instances of injuries are extremely rare, it is essential to maintain the temporal structure of the data when splitting it.

The approach used in this thesis is a grouped holdout approach where the data was grouped by athlete ID, ensuring that no data from the same athlete appeared in the training or validation sets. This approach preserves the temporal structure of the data and even though it is strict and may provide lower performance it reflects real-world deployment to unseen athletes better. Being able to adapt to unseen athletes was a higher priority since I believe such technology, that is able to mitigate harmful events, should be made available to a more general public.

Alternative strategies that could have been deployed are sliding window cross-validation or blocked time series validation where each fold respects the chronological order (Cerqueira et al. 2020). These approaches are worth noting as they would have been better if the goal was to forecast future periods for the same athletes that we have information about already.

2.5 Modeling Approaches

The modeling part of the thesis reflects my exploratory, trial-and-error process that was aimed at identifying modeling strategies that will be suitable for the temporal structure and class

imbalance of the injury prediction task. During the initial overview of the data, its temporal nature led me to think that a time-aware architecture would be the best approach. The XGBoost architecture adopted by Lövdal et al. (2021) made me hypothesize that because of its lack of inherent temporal modeling capabilities, perhaps deep learning approaches designed for time series data might perform better. This section examines the trade-offs of different models and architectures used and how well they capture the challenges of rare event prediction.

My first modeling effort focused on implementing a Long Short-Term Memory (LSTM) network, which seemed appropriate as it is widely used for its ability to capture long-term time dependencies (Hochreiter & Schmidhuber, 1997). LSTMs have been implemented in physiological monitoring and predictive maintenance, where the sequential signals evolve over time (Zhao et al., 2019; Johnson et al., 2016). In the context of injury forecasting, they seemed well suited to learning from the data that stretches over a 7-year period. However, in practice, the model underperformed and was difficult to tune. The sparsity of injury cases in the data made the model very prone to overfitting, a known issue for LSTM-based models in low-data regimes (Bzdok et al., 2018). These limitations led me to try different architectures capable of capturing the temporal dynamics with greater efficiency.

Following the exploration done with the LSTMs, a simpler recurrent model using Gated Recurrent Units (GRUs) was implemented. GRUs, although similar to the LSTMS, have a reduced number of parameters and a lower computational cost, which makes them suitable for smaller datasets (Cho et al., 2014). This architecture turned out to be more stable during training and slightly improved generalization compared to the LSTMs. However, the performance

remained suboptimal in the early epochs, where the model struggled to detect short-term temporal patterns effectively. This brought out a key limitation of GRUs: while effective for long-term dependencies, they do not perform as well at capturing sudden changes in training load or intensity. These findings were the reason for incorporating a Convolutional Neural Network (CNN) component to complement the GRU by extracting local features within the 20-day window.

The next architecture of the thesis is a deep learning hybrid model that consists of a 1D Convolutional Neural Network (CNN) followed by a Gated Recurrent Unit (GRU) layer. This architecture is selected because of the complementary strengths of convolutional and recurrent networks in modeling time series data, particularly in rare event prediction. The role of the CNN is to capture local short-term dynamics, for example, changes in training load, while the GRU is designed to capture long-term dependencies. Deep learning approaches are particularly useful for high dimensional time series data as they have automatic feature extraction and end-to-end training. Such CNN+RNN hybrids have been successfully applied in other domains for capturing rare events such as medical deterioration (Zhang et al., 2021) and industrial fault detection (Zhao et al., 2019). The hybrid model outperformed earlier LSTM and GRU-only architectures. Alongside the deep learning branch of the hybrid model, a pipeline computes a set of engineered statistical features over each 20-day window. These features include metrics such as mean, standard deviation, maximum, minimum, slope, and coefficient of variation for each raw feature. This strategy was adopted from previous literature showing that statistical summaries often outperform black-box models, especially in a low-signal setting (Claudino et al., 2019; Bzdok et

al., 2018). The statistical and deep branches are merged before feeding into a final dense layer for binary classification.

Finally, an approach similar to that of Lövdal et al. (2021) was taken, assessing whether ensemble methods could improve model robustness and a bagged version of the hybrid model was implemented. Bagging, short for bootstrap aggregating (Breiman, 1996), is an ensemble technique where multiple models are trained on different bootstrap samples of the training data and their results are aggregated afterward. In this thesis, the architecture used was the hybrid model, mentioned earlier, which combines a CNN+GRU branch with a dense branch processing engineered features. Five instances of the model were trained independently with different random seeds and varied training subsets which were grouped by athletes. In the inference part, the predicted probabilities of the models were averaged before implying the classification threshold.

The motivation to implement bagging was to reduce prediction variance and overfitting on the sparse injury labels and to increase the model sensitivity to borderline cases that the model might ignore. Ensemble methods are commonly used in imbalance learning scenarios where individual models struggle to generalize because of limited signals (Liu et al., 2009). Bagging comes with increased computational cost and its effectiveness depends on the diversity and the strength of the base learners which we will cover in the Results chapter.

2.6 Class Imbalances

When dealing with rare event prediction a common challenge will be dealing with an imbalanced dataset. In this thesis, the raw data set taken from Lövdal et al's (2021) study has an extremely imbalanced dataset where the injuries represent only 1% of the events in the data. With the data preprocessing and splitting techniques discussed earlier injuries account for around 15% of the total labeled sequences, varying based on the risk window selected. Without specific correction strategies machine learning algorithms tend to favor the majority class, which leads to a deceptively high accuracy but poor recall for the minority class. This is why addressing these imbalances is critical for performance and model reliability in real-world applications.

Several alternative techniques for handling the class imbalances were explored during the model development. One of them was Focal Loss, which is a modification of the standard cross-entropy loss that puts less weight on correctly classified examples and forces the model to focus more on hard misclassified cases (Lin et al., 2017). This is a great approach for rare event prediction like injury prevention but it requires careful tuning of its focusing and scaling hyperparameters. Other common resampling-based strategies such as SMOTE (Chawla et al. 2002) or undersampling the majority class were considered but as they introduce the risk of distorting the temporal patterns were not adopted.

The technique I found the most success with was class weighting where a higher penalty is assigned to misclassified positive samples in the loss function. Class weights are calculated inversely proportional to class frequencies and later integrated into the model's cross-entropy loss function. The use of class weights in neural network models is common as it is simple yet effective for tackling class imbalances (Elkan, 2001).

Additionally, threshold tuning was applied to improve the F1 score and balance between precision and recall. Switching from the default value of 0.5 allows the model to optimize its classification trade-off. The strategy is useful as evaluation metrics such as AUC may not fully capture the model's performance under class imbalance (Saito & Rehmsmeier, 2015). I will explore the different effects of threshold in the Results chapter.

2.7 Evaluation Metrics

Rare event prediction requires metrics that reflect your performance on the minority class rather than your overall accuracy as it can be misleading. This thesis focuses on precision, recall, F1-score, and precision-recall AUC (PR-AUC) over metrics like accuracy and ROC-AUC, as these metrics better capture the ability to detect injuries (Saito & Rehmsmeier, 2015). The balancing between precision and recall was evaluated by the F1-score which served as a summary metric. The evaluation was conducted on a held-out validation set, using classification tools from scikit-learn, and performance will be further explored in the following Results chapter.

3 Results

This chapter presents the results of the modeling experiments done to illustrate the methodological challenges of rare event prediction in time series data. The goal was to highlight the whole process from beginning to end and show how different modeling choices including architecture, feature engineering, loss functions, and class imbalance techniques affect the behavior and results of models trained on highly imbalanced data.

Across all models, predictive performance remained modest. Precision-recall AUC (PR-AUC) values ranged between 0.043 to 0.058 and F1-scores peaked at 0.135. These results fall short of those reported by Lövdal et al. (2021), who achieved significantly better results using a different modeling approach on the same dataset. This difference in performance is the main focus of this thesis: rare event prediction is highly sensitive to methodological details, and even though similar approaches were used in preprocessing splitting, they can lead to large variations in outcome.

The following section presents each model evaluated during the experimentation part. For each model, key performance metrics are reported and briefly interpreted, with attention brought to how they compare to the rest. In addition to comparing overall performance, the results also highlight the effects of different modeling choices like class imbalance handling, feature representation, and validation strategy. This analysis provides insight into which techniques helped address the main challenges of rare event prediction and which ones failed to produce meaningful results.

3.1 LSTM Model

The first model evaluated was a recurrent neural network based on Long Short-Term Memory (LSTM) units. It was trained on 30-day sliding windows of input features using binary cross-entropy loss and class weighting to address class imbalances. Following testing 30-day windows were selected as the optimal input window for this model and no engineered features were used as well. The validation strategy followed a grouped holdout approach, where no athlete ID was overlapping between training and validation sets.

The model achieved a precision-recall area under the curve (PR-AUC) of 0.055, which indicates a limited ability to discriminate between injury and non-injury classes in this highly imbalanced setting. The best F1-score achieved of 0.096 was obtained at a threshold of 0.50, where the precision was 0.065 and the recall was 0.184 (Figure 1).

These results indicated that the LSTM favors the majority class where it correctly identifies 4820 of 5429 non-injury samples but struggles to detect injuries with only 42 out of 228 true positives identified (Figure 1). The low precision also shows that the model has a high false positive rating where more than 90% of the positive predictions are incorrect.

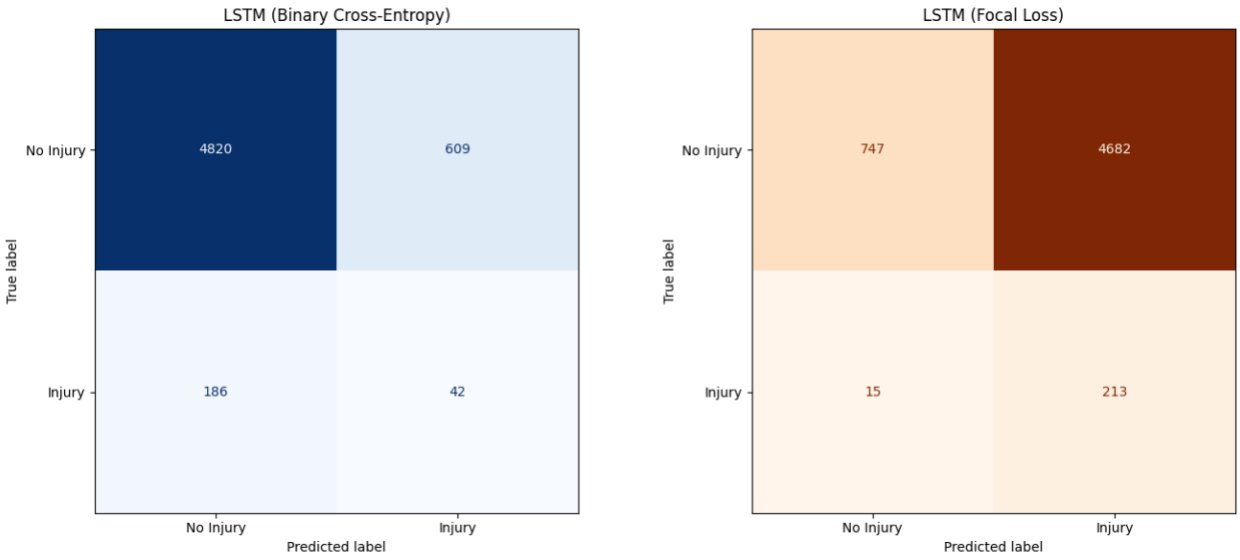

Figure 1: LSTM BCE and LSTM with Focal Loss Confusion Matrices

In order to see the impact of modifying the loss function for class imbalance, the LSTM architecture was re-trained using focal loss instead of the previously used binary cross entropy. Focal loss which is designed to focus training on hard-to-classify samples by down-weighting the well-classified samples fits well in this rare event context (Lin et al., 2017).

The LSTM model trained with focal loss achieved a PR-AUC of 0.046, which is lower than the standard LSTM. The best F1-score obtained was 0.083 with a notably high recall of 0.934 but a precision of just 0.044 (Figure 1).

The results show a drastic increase in recall compared to the previous loss function. However, this improvement comes at the cost of a very low precision, meaning that most of the predictions were false positives. This reflects a common pattern in highly imbalanced settings when using focal loss as it makes the model prioritize minority class detection at the expense of specificity.

3.2 GRU Model.

The next architecture was a different recurrent neural network, this time based on Gated Recurrent Units (GRUs). GRUs offer a simpler alternative to LSTMs, which often achieve similar performance with fewer parameters and faster training times (Chung et al., 2014). Unlike the LSTM model, the GRU was trained on 20-day time series windows using the same class weights and binary cross entropy loss.

The GRU model achieved a similar performance to that of the LSTM (Figure 4 and Figure 5) with a PR-AUC of 0.045 and an F1-score of 0.092 with precision of 0.058 and recall of 0.228(Figure 2).

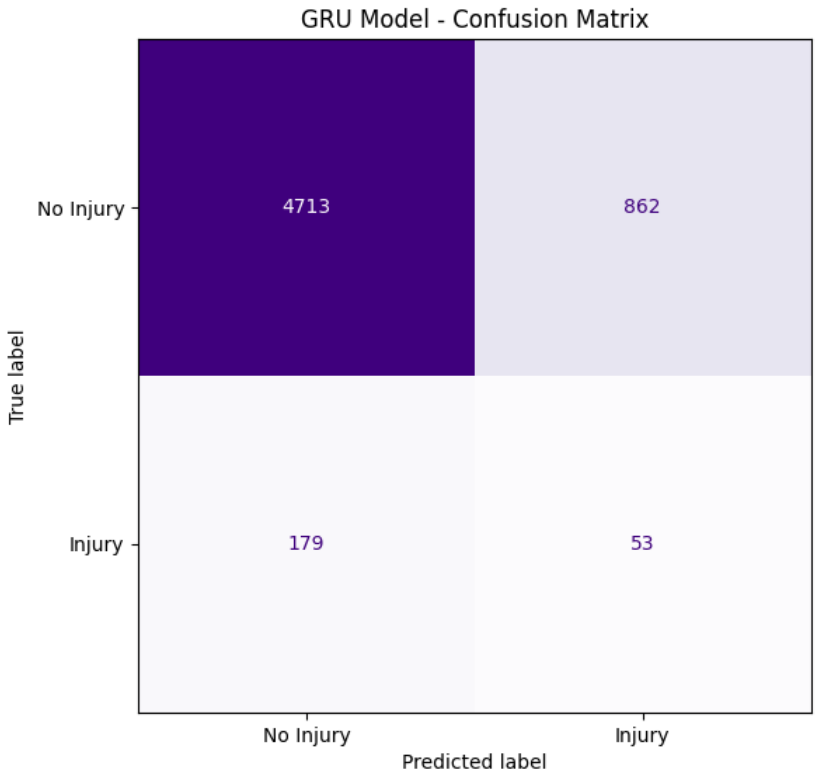

Figure 2: GRU Confusion Matrix

There were no significant differences across performance which suggests that GRUs may offer computational advantages but they do not offer a substantial performance boost under this setup.

3.3 Hybrid model

A hybrid model that combines deep learning approaches and engineered statistical summaries was engineered next. The deep learning part consisted of a convolutional neural network and a GRU applied to 30-day sequences. The output of both branches was combined and passed through a final dense layer for binary classification.

This hybrid model achieved the highest F1-score of all models tested - 0.135 with a precision of 0.093, recall of 0.246, and PR-AUC of 0.058 (Figure 3).

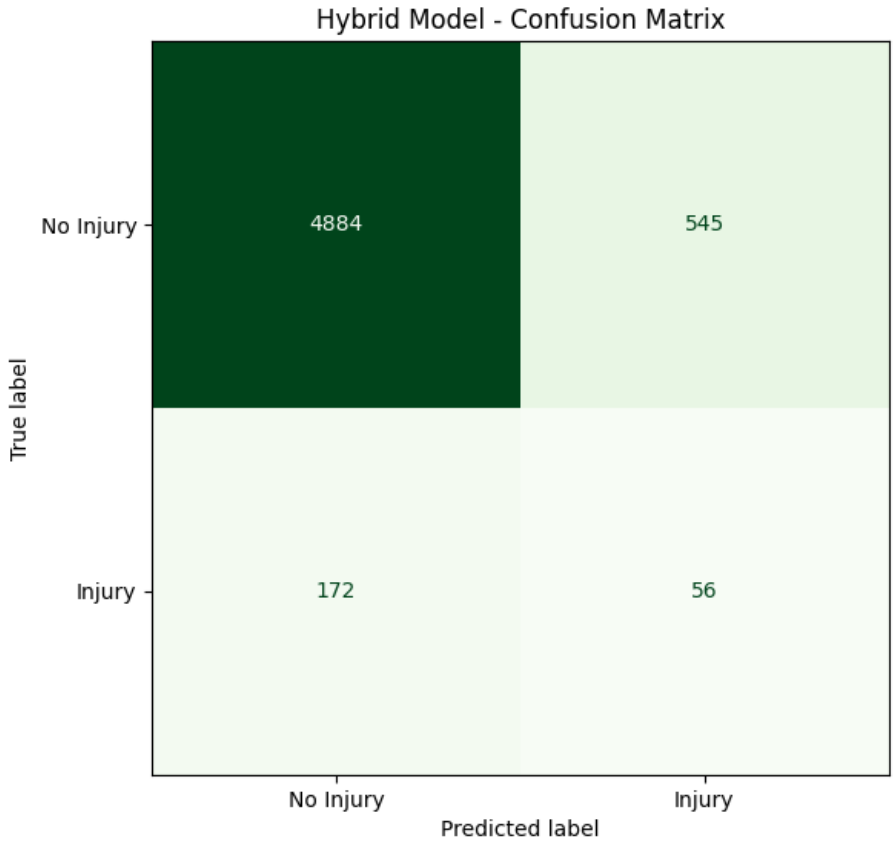

Figure 3: Confusion Matrix

The model was able to improve the precision of the previous models without having to sacrifice the recall. These results show that the addition of statistical features helps the model distinguish the injury class under extreme imbalance by summarizing the underlying trends that are missed by the deep learning models. While the performance is still poor it shows improvement from the previous models used.

3.4 Bagged Hybrid Model

As a final experiment, I applied a bagging (bootstrap aggregating) approach, similar to the one used in Lövdal et al.'s (2021) study. Bagging involves training multiple models on different random subsets of the data and averaging the results to reduce variance and overfitting. This method is often used to improve performance on noisy or high variance datasets (Breiman, 1996), and in the case of rare event prediction may help with prediction on the minority class. The bagged hybrid model was constructed by training 5 instances of the previously examined hybrid architecture.

The model achieved a PR-AUC of 0.043 and an F1-score of 0.080 with a precision of 0.055 and a recall of 0.211.

The bagged hybrid model achieved a better performance in terms of recall but everything else was lower compared to the standard hybrid model(Figure 4 and Figure 5). This suggests that the ensemble averaging increases the model's performance on hard cases thus increasing the true positives but it also introduces more false positives.

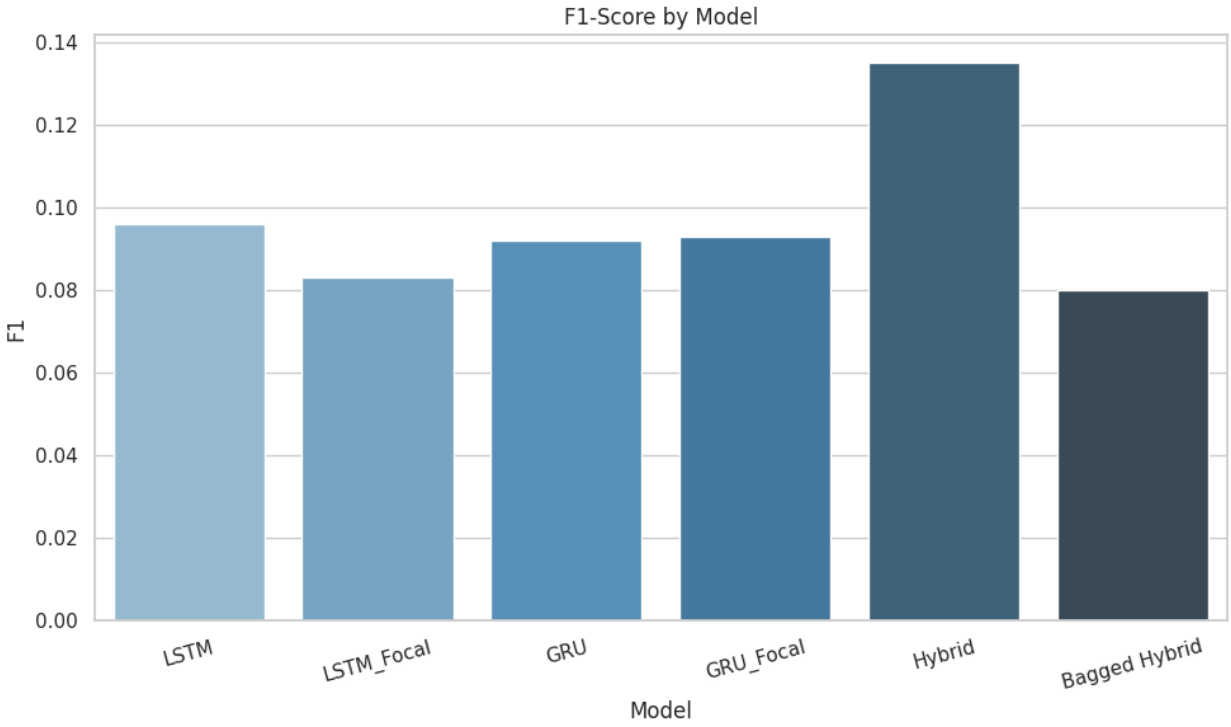

Figure 4: F1-Score by Model

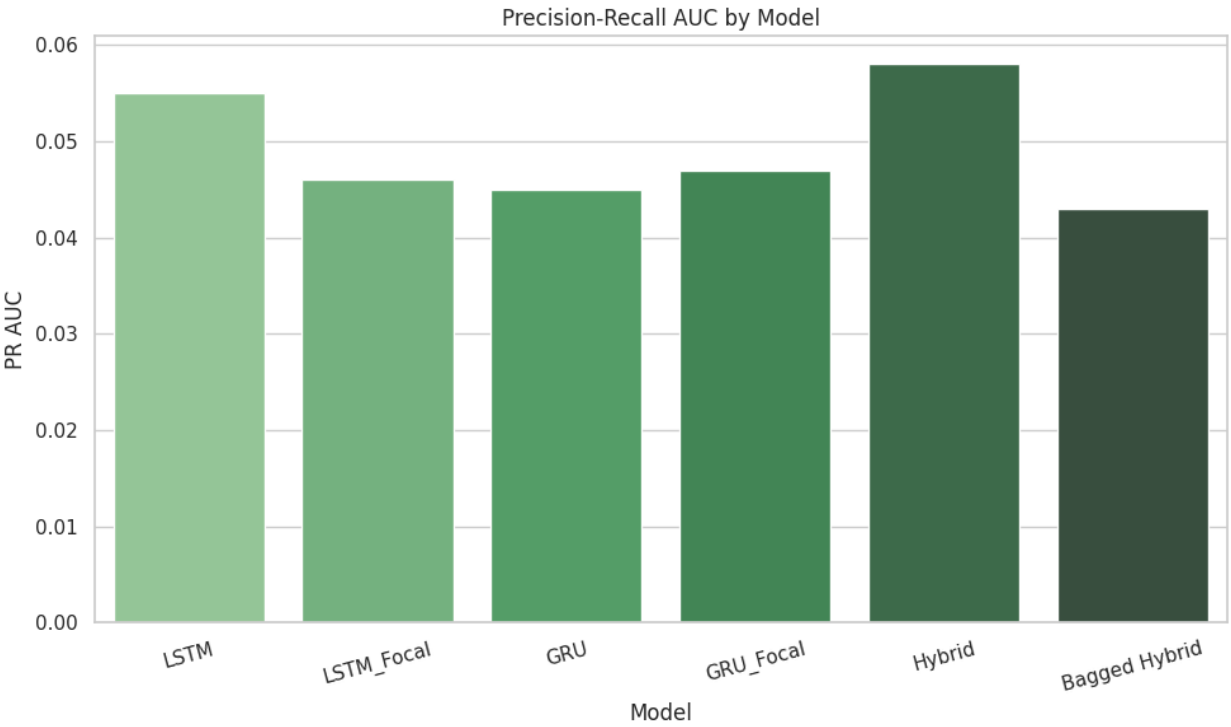

Figure 5: Precision-Recall AUC by Model

Across all model tests the performance remained modest. The best-performing model - the hybrid model combining statistical features with CNN+GRU layers achieved an F1-score of 0.135 and PR-AUC of 0.058. This small improvement over the rest of the models shows that a hybrid approach might be a viable solution for better performance.

Overall, these results reflect the difficulty of modeling rare events in time series, where even complex models struggle to generalize from sparse and noisy signals. They also reinforce the thesis' central point: methodological design decisions play a big part in the results of model architecture.

4 Discussion

4.1 Reflection

After quickly finding out that I won't be able to match the performance of the XGBoost architecture used in Lövdal et al.'s (2021) study and coming across the many challenges of rare event prediction in time series data, this thesis shifted focus from performance optimization to critical evaluation of the modeling pipeline (Haugen, 2017).

Despite the various models used and the implementation of hybrid architectures as well as multiple class imbalance and evaluation strategies the predictive performance I achieved remained modest. None of the models I tested achieved a significant PR-AUC or a reasonable F1-score, but these results are not just a modeling failure but rather evidence of the fragility and sensitivity inherent to rare event prediction in time series contexts.

One of the key reflections emerging from my modeling experiments is that architectural complexity on its own does not tackle the key challenges of rare event prediction. And while recurrent neural networks such as LSTMs and GRUs are theoretically very well-suited for the injury prediction case neither model was able to meaningfully distinguish between injury and non-injury sequences. The small performance gains observed when deploying the hybrid architecture show that incorporating engineered features helps compensate for the weak signal present in the raw sequences. This aligns with the findings of previous studies (Zhang et al. 2021, Claudino et al.,2019) that show the benefits of combining learned representations with hand-crafted ones in low-signal settings.

An interesting finding was that the use of ensembling did not provide the expected results. This is most likely due to the fact that ensembling weak learners who already struggle to find patterns will offer no gain. In my case, the individual hybrid models tended to make conservative predictions, and bagging such models amplified this conservatism leading to fewer positive predictions and lower recall.

These observations show that architectural choices matter but only when they are embedded within a well-designed learning pipeline. In rare event prediction, carefully designed input representations and well-handled class imbalances can be more valuable than deep or complex model structures alone.

Class imbalance, not surprisingly, proved to be one of the most difficult challenges throughout the modeling process. With injuries making up only a small percentage of the cases in the dataset, models often make conservative predictions achieving high accuracy while ignoring the minority class. Class weighting proved to be a good starting point which improved the performance slightly but it became clear that it was not enough to fight the big imbalance. Focal loss, which looked like a promising theory, offered disappointing performance in the end, highlighting the trade-off between sensitivity and reliability. The models tested with focal loss did not improve the general performance of the model compared to using class weights. These experiments emphasized that dealing with the imbalance is not about maximizing the metric you are chasing but finding the right balance for the domain. In rare event settings like injury

prediction, effective imbalance handling is very dependent on the context as well as the algorithms.

The findings of this thesis were focused on the specific case of injury prediction but they offer broader implications for rare event modeling in time series data. Different domains like healthcare, fraud detection, or equipment failure, share similar challenges - imbalanced data, noisy signals, and limited positive examples. The results suggest that in such context success often is the result of carefully aligning model design with domain constraints rather than deploying complex architectures. Ultimately this thesis shows that rare event prediction is not only a technical task but a methodological and domain-informed challenge, where the right combination of data framing, representation, and modeling strategy matter more than model complexity alone.

4.2 Limitations

This study faced several limitations. One of the main limitations was finding injury data for athletes. While Lövdal et al. 's (2021) dataset is one of the most comprehensive in the field; it reflects the broader data scarcity of the field. The availability of datasets including injuries is very low. This is most likely due to the fact that such data is tracked mainly by professional sports organizations that are not willing to share that data with their competition. Even though the Lovdal et al. (2021) dataset covers 74 athletes over 7 years, the features provided cover the on-track performance of the athletes, and no telemetric health data is provided. For a biomechanical field such as injuries such metrics would be really useful.

5 Conclusion

This thesis had the goal of exploring the question: What are the theoretical and methodological when trying to predict rare events in time series data? To investigate this a case study in sport injury prediction was done using the dataset introduced by Lövdal et al. (2021). Different models- from LSTMs and GRUs to hybrid deep learning architectures, were implemented and tested. Each model was evaluated using metrics focused on precision and recall under a grouped validation strategy that preserved the temporal and individual structure of the data.

The results revealed the complexity and fragility of rare event modeling. Despite the advanced architecture and techniques used to handle imbalances, the performance remained modest. The best model, a hybrid CNN+GRU combined with engineered features, achieved a low F1-score of just 0.135, with a PR-AUC of all models remaining below 0.06. The attempt to achieve results similar to those of Lövald et al. (2021) was unsuccessful, highlighting the difficulty of rare event time series tasks.

In other words, this work demonstrates that success in rare event prediction depends not only on the model but every component of the pipeline, including the preprocessing, feature engineering, and evaluation methodology. The importance of the domain in which the experiment is done is also highlighted as proper understanding of the setting is crucial for staying on the right track. The results support the idea that rare event prediction in time series is not only a technical problem but a methodological one, where robustness of results has greater priority over marginal performance gains.

Future research in rare event prediction and injury prediction, in particular, would benefit from methods that better handle noisy, limited, temporally structured data. One key area for improvement is the validation strategy. This thesis uses a held-out validation set; more flexible approaches such as stratified temporal sampling cross-validation could offer better balance and preserve the temporal structure, especially when datasets are small. What I would consider most critical for future work is an emphasis on reproducibility and methodological transparency. Many of the performance differences in this thesis come from small adjustments in preprocessing, labeling, or threshold tuning. Open end-to-end pipelines with documented design decisions would make comparisons more meaningful and interpretable in this field where small details drive large outcomes.

6 Appendix

https://github.com/mom4ilB/rare-event-injury-prediction

Bibliography

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling." *arXiv preprint arXiv:1803.01271*, 2018. https://arxiv.org/abs/1803.01271

Breiman, Leo. "Bagging Predictors." *Machine Learning* 24, no. 2 (1996): 123–140. https://doi.org/10.1007/BF00058655

Bzdok, Danilo, Naresh Subramaniam, and Bertrand Thirion. "The Contribution of Machine Learning to Neuroscience: Advances and Challenges." *Nature Neuroscience* 21, no. 11 (2018): 1549–1561. https://doi.org/10.1038/s41593-018-0262-9

Carey, David L., et al. "Risk Factors for Injury in Elite Australian Football Players: A Prospective Cohort Study." *Journal of Science and Medicine in Sport* 21, no. 3 (2018): 221–225. https://doi.org/10.1016/j.jsams.2017.06.006

Cerqueira, Vitor, et al. "Evaluating Time Series Classification." *Data Mining and Knowledge Discovery* 34, no. 3 (2020): 866–901. https://doi.org/10.1007/s10618-019-00659-5

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (2002): 321–357. https://doi.org/10.1613/jair.953

Cho, Kyunghyun, et al. "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation." *arXiv preprint arXiv:1406.1078*, 2014. https://arxiv.org/abs/1406.1078

Chung, Junyoung, et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." *arXiv preprint arXiv:1412.3555*, 2014. https://arxiv.org/abs/1412.3555

Claudino, João G., et al. "Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports." *Sports Medicine* 49, no. 9 (2019): 1441–1451. https://doi.org/10.1007/s40279-019-01129-4

Dal Pozzolo, Andrea, et al. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy." *IEEE Transactions on Neural Networks and Learning Systems* 29, no. 8 (2018): 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

Elkan, Charles. "The Foundations of Cost-Sensitive Learning." In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001. https://www.ijcai.org/Proceedings/01/Papers/032.pdf

Haugen, Thomas. "The Fragility of Injury Prediction Models: Statistical Power, Effect Sizes and Recommendations for Future Research." *Sports Medicine* 47, no. 9 (2017): 1795–1801. https://doi.org/10.1007/s40279-017-0671-7

He, Haibo, and Edwardo A. Garcia. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21, no. 9 (2009): 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." *Neural Computation* 9, no. 8 (1997): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Johnson, Alistair E.W., et al. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (2016): 160035. https://doi.org/10.1038/sdata.2016.35

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2015. https://mitpress.mit.edu/9780262029445

Krawczyk, Bartosz. "Learning from Imbalanced Data: Open Challenges and Future Directions." *Progress in Artificial Intelligence* 5, no. 4 (2016): 221–232. https://doi.org/10.1007/s13748-016-0094-0

Laine, Samuli, and Timo Aila. "Temporal Ensembling for Semi-Supervised Learning." *arXiv preprint arXiv:1610.02242*, 2017. https://arxiv.org/abs/1610.02242

Lin, Tsung-Yi, et al. "Focal Loss for Dense Object Detection." In *Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2980–2988. https://doi.org/10.1109/ICCV.2017.324

Liu, Xin Yao, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory Undersampling for Class-Imbalance Learning." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, no. 2 (2009): 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Lövdal, Torkel, et al. "Injury Prediction in Competitive Runners with Machine Learning." *Scientific Reports* 11, no. 1 (2021): 1–12. https://doi.org/10.1038/s41598-021-92070-4

Ramachand, Aditya. "Rare Event Prediction in Sequential Data Using Ensemble LSTMs." *arXiv preprint arXiv:2002.06513*, 2020. https://arxiv.org/abs/2002.06513

Rossi, Andrea, et al. "The Relationship between Training Load and Injury in Professional Soccer Players." *International Journal of Sports Physiology and Performance* 13, no. 5 (2018): 578–582. https://doi.org/10.1123/ijspp.2017-0336

Saito, Takaya, and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLoS ONE* 10, no. 3 (2015): e0118432. https://doi.org/10.1371/journal.pone.0118432

Shyalika, Chamathka, et al. "A Comprehensive Study on Rare Event Prediction." *Expert Systems with Applications* 234 (2024): 120222. https://doi.org/10.1016/j.eswa.2023.120222

Zhang, Yujie, et al. “Hybrid Modeling Approaches for Predicting Rare Events in Clinical Time Series Data.” *Journal of Biomedical Informatics* 118 (2021): 103778. https://doi.org/10.1016/j.jbi.2021.103778

Zhao, Rui, Ruqiang Yan, Zhen Lei, and Kai Mao. “Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks.” *IEEE Transactions on Industrial Electronics* 66, no. 10 (2019): 8653–8663. https://doi.org/10.1109/TIE.2018.2889774