WIRTSCHAFTSUNIVERSITÄT WIEN
Vienna University of Economics and Business

Bachelor's Thesis

Title of Bachelor's Thesis (English)	A Survey on Open Football Datasets and Their Utilization
Title of Bachelor's Thesis (German)	Eine Studie zu offenen Fußball-Datensätzen und deren Anwendungen
Author **(last name, first name):**	Nico Zandomeneghi
Student ID number:	11940024
Degree program:	Bachelor of Business and Economics, BSc (WU)
Examiner **(degree, first name, last name):**	Dr. Fajar J. Ekaputra

I hereby declare that:

1. I have written this Bachelor's thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.
2. This Bachelor's Thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.
3. This Bachelor's Thesis is identical with the thesis assessed by the examiner.
4. (Only applicable if the thesis was written by more than one author): this Bachelor's thesis was written together with

   The individual contributions of each writer as well as the co-written passages have been indicated.

25.09.2024
Date

Signiert von: Nico Zandomeneghi
Datum: 25.09.2024 20:11:27

A TRUST

Dieses Dokument ist digital signiert!
Dieses mit einer qualifizierten elektronischen Signatur versehene Dokument hat gemäß Art. 25 Abs. 2 der Verordnung (EU) Nr 910/2014 vom 23. Juli 2014 ("eIDAS-VO") die gleiche Rechtswirkung wie ein handschriftlich unterschriebenes Dokument.

Prüfinformation: Informationen zur Prüfung der elektronischen Signatur finden Sie unter: www.a-trust.at/pdf

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Bachelor Thesis

A Survey on Open Football Datasets and Their Utilization

Nico Zandomeneghi

Date of Birth: 05.06.2000
Student ID: 11940024

Subject Area: Information Business

Studienkennzahl: UJ 033 561

Supervisor: Dr. Fajar J. Ekaputra

Date of Submission: 25.09.2024

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK
UND OPERATIONS MANAGEMENT DEPARTMENT
OF INFORMATION SYSTEMS AND OPERATIONS
MANAGEMENT

Contents

List of Figures

List of Tables

Abstract

The integration of artificial intelligence (AI) and machine learning (ML) into football analytics has reshaped how player performance, recruitment, and tactical strategies are evaluated. However, the landscape of open football datasets, crucial for AI-driven analysis, remains underexplored. This thesis addresses the gap by systematically investigating openly accessible football datasets and evaluating their utility for AI applications. The research aims to assess the current state of these datasets, focusing on their quality, scope, and availability. A comprehensive survey of publicly available datasets was conducted, followed by the development of a common data model to map the datasets to typical analytical tasks. The study also explores the feasibility of automating dataset collection and identifies critical gaps between existing datasets and the data requirements for football analytics. Key findings highlight the presence of several comprehensive datasets covering top European leagues, but reveal gaps in advanced metrics, tactical analysis, and lower-tier league coverage. These limitations restrict the full potential of not only AI-driven analytics in football. The thesis concludes with recommendations for improving dataset accessibility and quality, emphasizing the need for standardized data structures to support future (AI) applications in sports analytics.

Keywords: football analytics, artificial intelligence, machine learning, open data, datasets, performance analysis, tactical analysis, data quality, model development

1 Introduction

The integration of artificial intelligence (AI) and machine learning (ML) techniques into sports analytics, particularly in football, has transformed how the sport is analyzed. AI-driven advancements in player selection, performance evaluation, and tactical strategy highlight the sport's increasing reliance on data-driven insights (Herberger and Litke, 2021). Despite these innovations, open football datasets remain largely underutilized, creating a research gap. As AI continues to evolve, the availability and quality of data are paramount for producing accurate and efficient analyses (Ćwiklinski et al., 2021).

The need for systematic documentation and assessment of these open football databases is therefore crucial. Football data is employed across multiple disciplines, from sports science, which uses it to track player metrics like speed and endurance (Prieto-Lage et al., 2021), to economics, where it provides insights into the financial dynamics of clubs and player transfers (Mrhari and Hasssouni, 2023b). Additionally, sociologists use football data to study fan behavior and the sport's cultural significance (Burton et al., 2019). However, while significant progress has been made in using AI for tasks like performance monitoring and team strategy, insufficient attention has been paid to the underlying datasets. This lack of focus on data quality and accessibility could limit the accuracy and effectiveness of AI-driven analyses (Owusu, 2008). This research aims to fill this gap by thoroughly evaluating the landscape of open football datasets, examining their characteristics and applications.

The central research question driving this thesis is:

- What is the current state of open football datasets and their utilizations?

This question is complemented by the following sub-research questions:

- RQ1: What types of openly accessible football datasets are available on the internet, and what are their attributes, including quality?
- RQ2: What kind of tasks are typically performed on football datasets?
- RQ3: What are the gaps between the available datasets and typical tasks?
- RQ4: To what extent can automation be employed in collecting open football datasets?

The methodology of this research involves a comprehensive survey of open football datasets, evaluating their completeness, accessibility, and relevance. A common data model is developed to map the datasets to typical analytical tasks, and a feasibility evaluation is conducted to assess the utility of these datasets in not only AI-driven tasks.

This thesis is structured as follows: first, a review of the existing literature on sports data analytics identifies research gaps in Chapter 2, particularly regarding the use of publicly available datasets. The methodology Chapter (Chapter 3) details the data collection and evaluation process, followed by an analysis of the results (Chapter 4), including the datasets discovered and their applicability. Finally, the thesis concludes with a discussion on the suitability of open datasets for various analytical tasks (Chapter 5) and provides recommendations for future improvements in data collection and availability (Chapter 6).

2 Related Work

The role of data in sports has evolved significantly over recent years, fundamentally changing how football, among other sports, is analyzed, managed, and understood. With the increasing availability of vast amounts of data and the growing sophistication of analytical techniques, football has become a prime field for applying data-driven insights (Watanabe et al., 2021). This shift has allowed for more informed decision-making both on and off the field, where performance metrics, predictive modeling, and advanced statistical analyses offer strategic advantages (Chandra et al., 2024). Data in football is no longer just about counting goals and assists; it now encompasses complex interactions such as player movements, match tactics, injury risks, and market value estimations (Benito Santos et al., 2018). The following sections explore the significance of data analytics in football, the tasks performed using datasets, the methodologies employed, and the automation of data collection, highlighting its multifaceted influence on the sport.

2.1 Importance of Data in Sports Analytics

Data analytics has revolutionized the world of sports, particularly football, by providing deep insights that enhance performance, strategy, and decision-making. The integration of sophisticated data analysis techniques has transformed football into a data-rich sport where every aspect of the game is quantified and analyzed. For instance, the utilization of big data and analytics in sport management has shown significant benefits in terms of improving team performance and fan engagement (Baumer and Zimbalist, 2014). The availability of comprehensive datasets enables detailed statistical analyses and the development of predictive models that can forecast player performance, game outcomes, and even injury risks. The use of data analytics in football is multifaceted, influencing both on-field and off-field activities. On the field, coaches and managers use data to devise game strategies, assess player performance, and make informed decisions during matches. Detailed statistical analysis can reveal patterns and trends that are not immediately apparent through traditional observation. For instance, advanced metrics can track player movements, ball possession, pass completion rates, and other critical aspects of gameplay, providing a comprehensive understanding of team dynamics (Benito Santos et al., 2018). Off the field, data analytics is crucial in player recruitment and market value estimation. By analyzing performance data, scouts and team managers can identify potential talent and make strategic decisions regarding player acquisitions. Predictive models can estimate a player's future performance based on historical data, helping teams invest wisely in new players. Additionally, data-driven insights are used to manage player health and fitness, predicting injury risks and optimizing training programs to enhance player longevity and performance (Chandra et al., 2024). The integration of Linked Open Data (LOD) in sports, particularly football, enhances the depth and breadth of available datasets, facilitating more sophisticated analyses and innovative applications that were previously limited by proprietary data constraints (Bergmann et al., 2013). This approach not only broadens the scope of data analysis but also promotes transparency and accessibility, allowing researchers and analysts to build upon existing data and develop new methodologies for sports analytics. Moreover, the use of data in sports extends beyond the field, influencing business decisions and market strategies within the football industry. A study highlighted that predictive analytics could significantly impact sports finance, such as assessing the impact of player injuries on the stock prices of football clubs using event study methodology and logistic regression (Mrhari and Hasssouni, 2023b). The fusion of data from various sources, including social media and weather data, enriches the analytical landscape, enabling comprehensive insights into both performance and external factors affecting the game (Bergmann et al., 2013). Data analytics also plays a crucial role in enhancing fan engagement. By analyzing fan behavior and preferences through social media and other digital platforms, clubs can tailor their marketing strategies to better engage with their audience. This not only improves fan satisfaction but also drives revenue through targeted promotions and personalized experiences (Watanabe et al., 2021).

2.2 Utilization and Tasks Performed on Football Datasets

Football datasets are widely used to perform various tasks that enhance both the strategic and operational aspects of the game. From predicting player performance and assessing tactical formations to estimating market values and preventing injuries, these datasets provide critical insights for decision-making. The following subsections explore the specific tasks performed on football datasets, the methodologies applied to analyze this data, and the automation processes involved in collecting it. Together, these aspects highlight the growing importance of data in modern football.

2.2.1 Overview of Tasks Conducted on Football Datasets

Football datasets are utilized for a myriad of tasks that significantly enhance the understanding and management of the game. Key tasks include player performance prediction, tactical analysis, injury prediction, and market value estimation. The dataset collected by Bergmann et al. (2013) from sources like UEFA.com and Fussballdaten.de includes extensive information on matches, players, teams, and in-game events, supporting various analytical tasks that are crucial for coaches and managers (Bergmann et al., 2013). Similarly, Liu et al. (2015) highlight the use of machine learning algorithms to predict player performance and analyze tactical formations, enabling teams to optimize their strategies based on data-driven insights (Liu et al., 2015). These tasks are crucial for optimizing player utilization and game strategies. For instance, injury prediction models help reduce player downtime by predicting potential injuries based on historical data and player health metrics (Mrhari and Hasssouni, 2023b). The application of machine learning techniques in predicting player performance has shown promising results, with models achieving significant accuracy in forecasting key performance indicators (Chandra et al., 2024). Tactical analysis is another critical area where data analytics is extensively used. By analyzing match data, teams can gain insights into the effectiveness of different formations and strategies. This information can be used to adjust tactics mid-game or prepare for future opponents by exploiting their weaknesses and countering their strengths. Advanced analytics also allow for the evaluation of individual player contributions to overall team performance, helping coaches make informed decisions about player selection and substitutions (Watanabe et al., 2021). Market value estimation is also significantly enhanced through data analytics. By analyzing various performance metrics and market trends, clubs can estimate the current and future market value of players. This information is invaluable during transfer windows, helping clubs make informed decisions about buying or selling players. Furthermore, data analytics can identify undervalued players, providing opportunities for strategic acquisitions that can strengthen the team without overspending (Payyappalli and Zhuang, 2019).

2.2.2 Methodologies Used for Data Analysis

Various methodologies are employed to analyze football datasets, each tailored to specific analytical needs. Machine learning algorithms, such as regression models and logistic regression, are commonly used for predictive tasks. For example, Mrhari and Hasssouni (2023b) utilized event study methodology and logistic regression to assess the impact of player injuries on the stock prices of football clubs, demonstrating the applicability of these methods in sports finance. Data visualization techniques also play a crucial role in presenting complex data in an accessible format, facilitating better decision-making. Visualization tools can help coaches and analysts quickly interpret data patterns and trends, making it easier to devise strategies and make informed decisions. Techniques such as heat maps, pass maps, and player movement charts are commonly used to visualize spatial data and player interactions on the field (Payyappalli and Zhuang, 2019). Additionally, advanced statistical methods and semantic technologies are used to integrate data from multiple sources, improving the overall quality and scope of analyses. Bergmann et al. (2013) describe how data from various sources are integrated using semantic technologies, enhancing the quality of the data and enabling more comprehensive analyses. Another study highlighted the use of sophisticated algorithms like Random Forest and Support Vector Machines (SVM) for classifying and predicting match outcomes, showcasing the versatility of machine learning in sports analytics (Chandra et al., 2024). Deep learning techniques are increasingly being adopted for more complex analyses, such as predicting match outcomes and player performance. These methods can handle large volumes of data and identify intricate patterns that traditional statistical methods might miss. For instance, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used to analyze sequential data, such as player movements and match sequences, providing deeper insights into game dynamics (Watanabe et al., 2021).

2.2.3 Automation of Dataset Collection

Automation in dataset collection is essential for maintaining the accuracy and timeliness of football data. Automated data collection involves using APIs and web crawlers to gather data from various online sources, such as sports websites, social media, and live match feeds. For example, the Linked Soccer Data project automates the collection of match results and player information from websites like UEFA.com and Fussballdaten.de, updating these datasets regularly during matchdays. This automation ensures that the datasets are current and comprehensive, which is crucial for accurate analysis and decision-making (Bergmann et al., 2013). With the rapid growth of information technology, the field of sports has seen an unprecedented increase in the availability and complexity of data. The integration of sophisticated data collection techniques has become imperative for leveraging the full potential of sports analytics. The automation of dataset collection in sports extends to various applications, notably in performance evaluation and prediction. Automated systems gather comprehensive data on athletes' historical performances, training behaviors, and psychological states. This data feeds into predictive models that utilize machine learning algorithms, such as neural networks and support vector machines, to forecast future performance and identify potential stars in sports. These applications not only enhance training and game strategies but also provide valuable insights for scouting and talent development (Bai and Bai, 2021). In addition to performance data, automated systems can collect and analyze environmental factors, such as weather conditions and playing surface quality, which can significantly impact match outcomes. By integrating this data with player and match statistics, teams can gain a holistic understanding of the factors influencing performance, allowing for more precise and effective decision-making (Bergmann et al., 2013).

Open Issues and Challenges. Despite the advancements, several challenges remain in the automation of sports data collection. Integrating disparate data sources into a unified platform is essential to overcome data silos and enable comprehensive analysis. Additionally, ensuring data privacy and protecting athletes' sensitive information is critical. Developing robust privacy protection mechanisms and fine-grained access control systems are necessary to address these concerns (Bai and Bai, 2021). Furthermore, the legal and ethical implications of using automated systems and big data in sports analytics require careful consideration to ensure compliance with regulations and the fair treatment of all stakeholders involved (Watanabe et al., 2021). Another significant challenge is the quality and reliability of the collected data. Automated systems must be meticulously designed to minimize errors and ensure data accuracy. This includes handling missing data, filtering out noise, and validating the authenticity of the sources. Moreover, the rapid evolution of technology and data sources requires continuous updates and adaptations of data collection systems to maintain their effectiveness and relevance (Payyappalli and Zhuang, 2019). Finally, the interpretability of complex models used in data analysis poses a challenge. While advanced machine learning models can provide highly accurate predictions, their complexity often makes it difficult to understand the underlying factors driving these predictions. Developing interpretable models and transparent analysis methods is crucial to ensure that insights derived from data analytics can be effectively communicated and utilized by coaches, players, and other stakeholders (Chandra et al., 2024).

3 Methodology

This chapter outlines the methodology employed to conduct a comprehensive survey of open football datasets and their utilisation. The primary objective is to document and analyse existing datasets in order to gain an understanding of their characteristics, quality, and availability. This methodology is divided into several key components: data collection, data model development, data mapping, and feasibility evaluation. Each section details the processes, techniques, and criteria used to achieve the research objectives.

3.1 Data Collection

This Section details the methodology and metrics used for data collection on existing open football dataset and their licensing.

3.1.1 Survey of Existing Open Football Datasets

The survey began by identifying a range of open football datasets available on the web. This process involved extensive online searches and reviews of academic publications. Each dataset was evaluated based on its accessibility, comprehensiveness, and the relevance of its data to current football analytics practices. Various search engines and databases were utilized to gather the most up-to-date and publicly accessible datasets. Furthermore, the review process included assessing the quality of the data by examining factors such as the frequency of updates, the level of detail in player and match statistics, and the presence of historical data which could enable trend analysis. Forums and online communities dedicated to sports analytics were also scoured to gauge the usability of different datasets. This approach was aimed at ensuring that the datasets considered for further analysis would meet the high standards necessary for robust football analytics.

3.1.1.1 Completeness of Metadata The completeness of metadata was evaluated based on several comprehensive criteria, ensuring a thorough understanding and proper usage of the dataset. These criteria are detailed as follows:

Dataset Description. The dataset description provides crucial insight into the overall purpose, scope, and key features of each dataset. A clear and detailed description is vital for understanding the datasets content and how it aligns with the objectives of football analytics. In particular, it helps users identify whether the dataset is suitable for their specific analytical needs.

- **Detailed Overview of the Dataset:** A comprehensive summary describing the dataset, including its purpose, scope, and key features. This should provide a clear understanding of what the dataset encompasses and its primary objectives.

  - Excellent: Comprehensive summary with clear purpose, scope, key features.
  - Good: Good summary with most aspects covered.
  - Adequate: Basic summary with essential information.
  - Fair: Partial summary with significant missing details.
  - Poor: No summary provided.

- **Data Source:** Information about the origin of the dataset, including the name of the organization or individuals who collected or generated the data.
  - Excellent: Clear information about the origin, including names of collectors or generators.
  - Good: Mention of the source with some details.
  - Adequate: Basic mention of the source.
  - Fair: Partial information about the source.
  - Poor: No information about the source.

Variables and Definitions. The inclusion of a comprehensive list of variables and definitions is critical for the accurate interpretation of data. Each dataset was examined to ensure that all relevant variables were clearly enumerated, with definitions provided for each. This helps users avoid ambiguity and ensures that the methodology used to collect or measure the variables is well understood.

- **Comprehensive List of Data Included:** A comprehensive enumeration of all variables present in the dataset. Each variable should be accompanied by clear definitions and explanations, ensuring that users are able to comprehend the meaning of each variable and the methodology employed to collect or measure it.
  - Excellent: Complete enumeration with clear definitions and explanations.
  - Good: Comprehensive list with minor details missing.
  - Adequate: Basic list with essential variables.
  - Fair: Partial list with significant gaps.
  - Poor: No list or definitions provided.

Update Schedules. Datasets that are regularly updated tend to be more reliable for dynamic, ongoing analysis. Therefore, the presence of update schedules and the date of the most recent update is assessed to gauge the currency of the data. Knowing how often the dataset is updated, and when it was last refreshed, is essential for understanding its relevance to contemporary studies and ensuring its applicability to real-time or historical analyses.

- **Information on Update Frequency:** Details on how regularly the dataset is updated, including specific intervals (e.g., daily, monthly, annually). This helps users understand how current the data is and plan for future data updates.
  - Excellent: Detailed update schedule with specific intervals.
  - Good: Mention of update frequency with some details.
  - Adequate: Basic mention of updates.
  - Fair: Partial information on updates.
  - Poor: No information on update frequency.
- **Date of Most Recent Update:** The exact date when the dataset was last updated. This provides critical information on the currency and potential relevance of the data at the time of its usage.
  - Excellent: Exact date provided.
  - Good: Recent date mentioned but not exact.
  - Adequate: Basic mention of recent updates.
  - Fair: Partial information on the date.
  - Poor: No information on the most recent update.

3.1.1.2 Content Analysis

The content analysis was conducted to evaluate the comprehensiveness and depth of the dataset based on several specific criteria. These criteria are detailed as follows:

Data Coverage. In terms of data coverage, the geographical regions and seasons covered by the dataset were key factors in determining its scope. Datasets that include a wide range of regions, leagues, or countries, and cover multiple seasons, provide a broader, more representative picture of football activity, which is crucial for longitudinal analyses and comparisons across different footballing contexts.

- **Geographical Regions Covered:** An assessment of the geographical scope of the dataset, detailing the specific regions, leagues, or countries included. This criterion evaluates the breadth of the dataset in terms of its geographical inclusivity, ensuring it covers a representative and relevant set of regions pertinent to the study.
  - Excellent: Detailed geographical scope including specific regions, leagues, or countries.
  - Good: Comprehensive mention of regions.
  - Adequate: Basic mention of some regions.
  - Fair: Partial information on regions.
  - Poor: No information on geographical coverage.
- **Seasons Covered:** Information on the temporal range of the data, including specific years or range of years for which the data is available. This ensures the dataset provides a comprehensive longitudinal view necessary for trend analysis and longitudinal studies.
  - Excellent: Detailed temporal range including specific years or range of years.
  - Good: Comprehensive mention of seasons.
  - Adequate: Basic mention of some seasons.
  - Fair: Partial information on seasons.
  - Poor: No information on seasons covered.

Data Granularity. Another important aspect is the data granularity. The level of detail, whether at the match-level, player-level, or event-level, determines the extent to which fine-grained analysis can be conducted.

- **Level of Detail Provided:** An evaluation of the granularity of the data, examining whether the data is provided at a match-level, player-level, or event-level. This criterion assesses how detailed and specific the data is, which is crucial for analyses requiring fine-grained data points.
    - Excellent: Detailed granularity such as match-level, player-level, or event-level.
    - Good: Good level of detail with minor gaps.
    - Adequate: Basic level of detail.
    - Fair: Partial detail with significant gaps.
    - Poor: No detailed data provided.

Types of Data Included. The types of data included is also considered. A diverse dataset allows for more extensive analysis and offers greater potential for cross-referencing different types of information, thus enriching the research process.

- **The completeness of data types coverage:** A detailed examination of the different types of data present in the dataset. This includes, but is not limited to, match results, player statistics, and event data. This criterion ensures a comprehensive understanding of the various data dimensions available for analysis and their potential applications in the study.
    - Excellent: Detailed examination of various types of data (e.g., match results, player statistics, event data).
    - Good: Comprehensive types of data with minor details missing.
    - Adequate: Basic mention of types of data.
    - Fair: Partial information on types of data.
    - Poor: No information on types of data.

3.1.1.3 Accessibility

Accessibility of the datasets was evaluated based on several key criteria to ensure ease of access and usability. These criteria are detailed as follows:

Download Formats. The accessibility of the datasets was evaluated based on the available download formats.

- **Available Formats:** An assessment of the various formats in which the datasets can be downloaded. Common formats include CSV, JSON, and XML. This criterion evaluates the flexibility and convenience of accessing the data in different formats to suit various analytical tools and user preferences.
  - Excellent: Available in multiple formats (e.g., CSV, JSON, XML).
  - Good: Available in a few formats.
  - Adequate: Available in one format.
  - Fair: Available in a format but with issues.
  - Poor: Not specified or difficult to access.

3.1.2 Licensing Restrictions Methodology

The evaluation of licensing restrictions for the datasets was carried out to ensure that users are aware of the legal and ethical boundaries of using the datasets. The methodology involved assessing several key aspects, including the types of licenses, conditions, limitations, and usage permissions.

3.1.2.1 Licensing Conditions Licenses often come with specific conditions that dictate how the data can be used, modified, and distributed. The conditions evaluated in this study include:

- **License and copyright notice:** Requirement to include the original license and copyright notice.
- **Attribute:** Requirement to give appropriate credit to the original creator.
- **Share-Alike:** Requirement that derivatives must be licensed under identical terms.
- **Keep open:** Requirement to keep the data open and freely accessible.
- **Copyleft:** Requirement to allow distribution of derivative works under the same terms.
- **Disclose source:** Requirement to make the source code available if the data is modified.
- **Same license:** Requirement that derivatives must carry the same license.

- **State changes:** Requirement to indicate any changes made to the original data.
- **Network use is distribution:** Requirement that network use of the data counts as distribution.

These conditions ensure that the datasets are used in a manner that respects the original creators intentions and maintains the integrity and openness of the data.

3.1.2.2 Licensing Limitations Licenses may also include limitations that restrict certain uses of the data. The limitations assessed in this study include:

- **No Liability:** The data provider is not liable for any damages resulting from the use of the data.
- **No Warranty:** The data is provided without any warranty, including fitness for a particular purpose.
- **Sublicense:** Restriction on the ability to sublicense the data to third parties.
- **Trademark use:** Restriction on the use of trademarks associated with the data.

Understanding these limitations helps users to be aware of what they cannot do with the data, thereby avoiding legal issues and misuse.

3.1.2.3 Usage Permissions The study also evaluated the permissions associated with each license type to determine the scope of allowable actions:

- **Private Use:** Whether the dataset can be used for personal, non-commercial purposes.
- **Commercial Use:** Whether the dataset can be used for commercial purposes, allowing the user to profit from the data.
- **Modify:** Whether the dataset can be altered, transformed, or built upon.
- **Distribute:** Whether the dataset can be shared or redistributed, either in its original form or as part of derivative works.
- **Patent Use:** For some licenses, there may be specific permissions or restrictions related to the use of patents. This includes whether the dataset can be used in ways that involve patented technologies or innovations.

3.2 Survey of Football Dataset Analysis Tasks

To gain a comprehensive understanding of the typical tasks performed on football datasets, an extensive and detailed survey of common analysis tasks was conducted. This involved thorough research across various resources, including academic papers, online forums, and specialized websites dedicated to sports results.

These tasks were categorized into major groups, with specific methodologies and tools employed for each task carefully documented. The survey aimed to capture not only the breadth but also the depth of analytical practices within the domain of football data analysis. For more detailed descriptions of the specific tasks and their associated methodologies and tools, refer to Appendix D.

3.3 Data Model Development

To ensure a consistent and standardized approach to handling football data, a common data model was developed. This model was designed to integrate multiple open datasets, serving as a proof of concept to demonstrate the feasibility of joining datasets to expand the range of analytical tasks that can be performed.

Common Data Model for Open Football Datasets: The common data model was developed using a relational database approach, implemented in PostgreSQL. This choice leverages PostgreSQL's robust capabilities for handling complex queries, ensuring data integrity, and providing scalability. The development process involved selecting a primary open dataset and incrementally integrating additional datasets to enhance the analytical capabilities.

3.4 Mapping Datasets and Tasks to the Common Data Model

The process of mapping datasets and analytical tasks to the common data model involves several structured steps to ensure the integrated usability and completeness of the data:

Creating a Unified Data Schema. The most comprehensive dataset schema was selected as the base. Each additional dataset was mapped to this schema, extending it to incorporate unique variables from each dataset. This iterative process ensured that the final schema encompassed the full range of available data.

Descriptive Table Creation For each dataset, a detailed table was created that described which leagues and types of data (e.g., player statistics, match results) were included. This table served as a reference for mapping datasets to the unified schema.

Schema Extension and Mapping The data schema was extended incrementally. For each new dataset, variables were aligned with the existing schema, and new entries were added where necessary. This involved:

- **Identifying Corresponding Variables:** Matching variables from the new dataset to existing schema variables.
- **Handling Discrepancies:** Applying transformations and normalizations to ensure compatibility.
- **Integrating Unique Data:** Adding new variables to the schema to capture unique data points from each dataset.

Mapping Analytical Tasks Common analytical tasks were identified and categorized. Each task was then mapped to the unified schema to determine the data requirements. This step involved:

- **Task Definition:** Documenting typical tasks (e.g., player performance analysis, match outcome predictions) performed on football datasets.
- **Data Requirements Assessment:** Determining the specific data needed for each task and verifying its availability within the integrated datasets.
- **Traceability:** Establishing a traceability matrix that linked each task to the relevant data sources, allowing for the identification of necessary datasets for each analysis.

3.5 Feasibility Evaluation of Datasets for Analytic Tasks

The feasibility of using the integrated datasets for various analytic tasks was evaluated based on several quality criteria:

- **Data Size**: The size of the dataset in terms of the number of records and data points was assessed to determine if it is sufficiently large to support robust statistical analysis and machine learning applications. Larger datasets generally provide more comprehensive insights but may require more substantial computational resources.
- **Frequency of Updates**: The frequency with which the dataset is updated with new information was evaluated. Frequent updates ensure that the data remains current and relevant for ongoing analyses, enhancing the timeliness of the data.

- **Documentation**: The availability and quality of documentation that describes the dataset and its variables were assessed. Comprehensive documentation, including descriptions of variables, data collection methods, and any preprocessing steps, is crucial for understanding the context and limitations of the data.

- **Usability of Data Structure**: The structure of the data and its ease of integration and use for analysis were evaluated. This involved checking for a well-defined schema, clear relationships between data tables, and consistent data types, ensuring that the data is conducive to analysis.

- **Time Range**: The temporal coverage of the dataset, i.e., the range of dates for which data is available, was assessed to ensure it covers a sufficiently long time period for longitudinal analyses and trend identification.

4 Results

This Chapter reports on the result of our research, which we categorized into five sections: (i) Discovery of Open Football Datasets, (ii) Categorization of Analytical Tasks, (iii) Common Data Model, (iv) Mapping Datasets to Analytical Tasks, and (v) Feasibility Evaluation.

4.1 Discovery of Open Football Datasets

The search process yielded a total of 21 publicly available football datasets, of which eight met the requisite criteria for accessibility and licensing. These datasets varied in size, scope. focus, offering a all a different range of data. In Table 1 datasets which have a license where data is open to use are marked as open in the accessibility column. Data that is open but also, in a sense, commercially closed - for instance, free but then protected after a certain amount of requests - is designated as semi-open, while data lacking any license information whatsoever is classified as gray. The two aforementioned categories will not be further discussed in this thesis as this datasets are deemed to be of limited relevance to the subject matter.

4.1.1 Licenses

It is important to note that even when datasets are openly accessible, they may still be subject to different open licenses. In this case, six out of our eight datasets have different licenses. These licenses may be employed for both private and commercial purposes, modified, and distributed. Nevertheless, they are subject to disparate conditions and limitations.

Conditions:

- Condition 1: Requires inclusion of the license and copyright notice to ensure acknowledgment of the original source.
- Condition 2: Mandates attribution to the original creator, promoting transparency and giving proper credit.
- Condition 3: Share-alike or copyleft provisions require that any derivative works be distributed under the same terms, influencing how new research is shared.
- Condition 5: Copyleft provisions may limit integration with differently licensed projects.

Dataset	**Accessibility**	**License**
European Soccer Database	open	ODbL v1.0
Football (Soccer) Data for Everyone	open	MIT License
football.db	open	CC0 (public domain)
International football results from 1872 to 2024	open	CC0 (public domain)
OpenLigaDB	open	Apache License 2.0
Transfermarkt Datasets	open	CC0 (public domain)
WoSo Stats	open	GPL-3.0
Worldcup	open	CC BY-SA 4.0 LEGAL CODE
API Football	semi-open	none
Football-data org	semi-open	none
RapidAPI (API-Football)	semi-open	none
Soccer Video Api	semi-open	none
UCL Stats	gray	none
Football Data.co.uk	gray	none
BetGPS	gray	none
Connect API	gray	none
English and European soccer results 1871-2022	gray	none
football-standings-api	gray	none
Long time game data (all leagues)	gray	none
Squiggle api (AFL)	gray	none
The SportsDB	gray	none

Table 1: Dataset Accessibility and License Information

- Condition 6: Requires disclosure of the source, promoting transparency and documentation.
- Condition 8: Requires stating changes made to the original work, ensuring openness in modifications.

Limitations:

- Limitation a: No liability, shifting responsibility for the data's use and reliability to the user.
- Limitation b: No warranty, further emphasizing that users are responsible for assessing the data's reliability.
- Limitation c: Restrictions on sublicensing may affect how the dataset can be redistributed.
- Limitation d: Trademark use may introduce additional legal considerations that can affect branding and redistribution.

Understanding these varied conditions and limitations is essential for researchers and practitioners to ensure compliance and promote responsible use of open data within the scientific community.

Licence	**Private Use**	**Commercial Use**	**Modify**	**Distribute**	**Conditions**	**Limitation**	**Patent Use**
CC0 (public domain)	x	x	x	x	1	a	
MIT	x	x	x	x	[1,2]	[a,b]	
Open Data Commons Open Database License (ODbL) v1.0	x	x	x	x	[1,2,3,4,5]	[a,b,c]	
GNU General Public License	x	x	x	x	[1,6,7,8,9]	[a,b]	x
Apache License 2.0	x	x	x	x	[1,8]	[a,b,d]	x
CC BY-SA 4.0 LEGAL CODE	x	x	x	x	3	c	

Table 2: License Information Table (LIT)

Reference	**Condition**
1	License and copyright notice
2	Attribute
3	Share-Alike
4	Keep open
5	Copyleft
6	Disclose source
7	Same licence
8	State changes
9	Network use is distribution

Table 3: LIT: Conditions

Reference	**Limitation**
a	No Liability
b	No Warranty
c	Sublicense
d	Trademark use

Table 4: LIT: Limitations

4.1.2 Summary of Open Football Datasets

The presented tables offer a comprehensive overview of various open football datasets and their coverage of different football competitions, both domestic leagues and cups, as well as international tournaments.

League Competitions: The Table 5 focuses on the coverage of top-tier European league competitions across the datasets. The leagues considered are the Premier League (England), Bundesliga 1 (Germany), Serie A (Italy), La Liga (Spain), and Ligue 1 (France). The datasets **European Soccer Database**, **Football (Soccer) Data for Everyone**, **football.db**, **OpenLigaDB**, and **Transfermarkt Datasets** all provide data for these five leagues, indicating a comprehensive coverage of major European football leagues. In contrast, the datasets **International football results from 1872 to 2024**, **WoSo Stats**, and **Worldcup** do not offer data on these league competitions.

	Premiere Leage	**Bundesliga 1**	**Serie A**	**La Liga**	**Ligue 1**
European Soccer Database	x	x	x	x	x
Football (Soccer) Data for Everyone	x	x	x	x	x
football.db	x	x	x	x	x
International football results from 1872 to 2024					
OpenLigaDB	x	x	x	x	x
Transfermarkt Datasets	x	x	x	x	x
WoSo Stats					
Worldcup					

Table 5: Data sources: league competitions.

Cup Competitions: The Table 6 outlines the availability of data for key domestic cup competitions, namely the FA Cup (England), DFB-Pokal (Germany), Coppa Italia (Italy), Copa del Rey (Spain), and Coupe de France (France). The **football.db** and **Transfermarkt Datasets** cover all listed cup competitions, suggesting extensive data on national cup tournaments. **OpenLigaDB** provides data for the FA Cup and DFB-Pokal but lacks coverage for the Coppa Italia, Copa del Rey, and Coupe de France. The other datasets, including **European Soccer Database**, **Football (Soccer) Data for Everyone**, **International football results from 1872 to 2024**, **WoSo Stats**, and **Worldcup**, do not include data on these cup competitions.

	FA CUP	**DFB-Pokal**	**Coppa Italia**	**Copa del Rey**	**Coupe de France**
European Soccer Database					
Football (Soccer) Data for Everyone					
football.db	x	x	x	x	x
International football results from 1872 to 2024					
OpenLigaDB	x	x			
Transfermarkt Datasets	x	x	x	x	x
WoSo Stats					
Worldcup					

Table 6: Data sources and cup competitions.

International Competitions: The Table 7 presents data coverage for major international competitions such as the UEFA Champions League, UEFA Europa League, UEFA Europa Conference League, FIFA World Cup, and UEFA European Championship. The **Transfermarkt Datasets** encompass all these international tournaments, reflecting a wide-ranging dataset for international football. **football.db** includes data on the UEFA Champions League and FIFA World Cup, while **International football results from 1872 to 2024** covers the UEFA Champions League, FIFA World Cup, and UEFA European Championship. **OpenLigaDB** provides information on the UEFA Champions League, UEFA Europa League, and UEFA Europa Conference League but does not cover the FIFA World Cup or UEFA European Championship. The **WoSo Stats** dataset includes data for the FIFA World Cup and UEFA European Championship, indicating a focus on international competitions. The **Worldcup** dataset is specialized, providing data exclusively on the FIFA World Cup. The **European Soccer Database** and **Football (Soccer) Data for Everyone** do not offer data on these international competitions.

	UEFA Champions League	UEFA Europa League	UEFA Conference League	FIFA World Cup	UEFA European Championship
European Soccer Database					
Football (Soccer) Data for Everyone					
football.db	x			x	
International football results from 1872 to 2024	x			x	x
OpenLigaDB	x	x	x		
Transfermarkt Datasets	x	x	x	x	x
WoSo Stats				x	x
Worldcup				x	

Table 7: Data sources and international competitions.

While the tables provide an overview of the competitions covered by each dataset, the specific years of data available vary between datasets. Some datasets may offer extensive historical data spanning many decades, while others focus on more recent seasons or specific time periods.

4.1.3 Dataset Quality

The dataset quality was evaluated across ten criteria ranking them by numbers from 1 to 5 (1 = excellent, 5 = poor). This grading system has been applied to the open football datasets to assess their usability, comprehensiveness, and accessibility for football analytics.

Dataset Description The Detailed Overview of the Dataset evaluates how well the dataset's scope and key features are summarized. Higher-scoring datasets such as **football.db** (1) provided clear summaries, while **European Soccer Database**

Criteria	**Dataset Sources**							
	European Soccer Database	Football (Soccer) Data for Everyone	football.db	International football results from 1872 to 2024	OpenLigaDB	Transfermarkt Datasets	WoSo Stats	Worldcup
Detailed Overview of the Dataset	3	3	1	1	1	1	2	1
Data Source	1	1	5	4	3	3	5	2
Comprehensive List of Data Included	1	5	3	1	5	1	4	1
Information on Update Frequency	5	5	5	1	5	1	4	5
Date of Most Recent Update	5	5	5	2	5	2	5	5
Geographical Regions Covered	1	1	1	1	3	1	5	1
Seasons Covered	1	1	5	1	5	2	3	1
Level of Detail Provided	1	1	3	3	4	2	3	3
Types of Data Included	1	3	3	3	3	2	3	3
Available Formats	3	2	3	3	3	2	3	1
Average Points	2.2	2.7	3.4	2	3.7	1.7	3.7	2.3

Table 8: Comparison of Various Football Datasets

(3) and **Football (Soccer) Data for Everyone** (3) offered less detailed summaries, impacting their usability.

Variables and Definitions The Comprehensive List of Data Included refers to how complete the enumeration of variables is. **OpenLigaDB** and **Football (Soccer) Data for Everyone** received lower grades (5), indicating highly comprehensive datasets. On the other hand, **Transfermarkt Datasets** (1) and **International football results from 1872 to 2024** (1) lacked in this area, affecting their utility for in-depth research.

Update Schedules The Information on Update Frequency and Date of Most Recent Update were crucial in determining how current each dataset is. Several datasets, including **football.db** and **OpenLigaDB** (both 5), performed poorly here, meaning they are frequently updated. **International football results from 1872 to 2024** had only occasional updates (1), and thus, may not always contain the latest information.

Data Coverage Geographical Regions Covered and Seasons Covered were vital for evaluating the breadth of the datasets. **European Soccer Database**, **Football (Soccer) Data for Everyone,** and **football.db** (all 1) excelled in these categories, as they cover a wide range of regions and seasons. In contrast, **WoSo Stats** (5) had limited geographical scope, which restricts its broader applicability.

Data Granularity The Level of Detail Provided shows how granular the data is (match-level, player-level, etc.). **OpenLigaDB** (4) had the most detailed data, which is beneficial for comprehensive football analysis. Other datasets were adequate, but some lacked the granularity necessary for in-depth studies.

Types of Data Included The Types of Data Included evaluates whether datasets include essential football data like match results and player statistics. Here, datasets like **European Soccer Database** (1) and **football.db** (3) performed well, whereas others lacked certain types of data.

Accessibility Finally, the Available Formats criterion assessed how easily the datasets could be accessed in various formats like CSV, JSON, or XML. **Worldcup** (1) provided excellent accessibility, while others like **Transfermarkt Datasets** (2) and **football.db** (3) offered fewer format options, limiting flexibility.

4.2 Categorization of Analytical Tasks

The common analysis tasks were categorized into several major groups, each focusing on different aspects of football performance and management:

Player Performance Analysis: This category focuses on evaluating individual player contributions by analyzing their playing time, offensive impact through goals and assists, and their disciplinary records, which affect availability and long-term performance.

– **Games/Minutes Played:** Analysis of player participation metrics to evaluate playing time and its correlation with performance metrics.

– **Scores and Assists:** Detailed examination of goal-scoring and assist-providing patterns to assess offensive contributions.

– **Disciplinary Records:** Analysis of cards (yellow and red) received to understand disciplinary issues and their impact on player availability.

Team Performance Analysis: Team-level metrics are examined here, aggregating key statistics such as win/loss ratios and advanced metrics like Expected Goals (xG). The aim is to provide insights into overall team efficiency and game outcomes.

– **Calculation of Team Statistics:** Aggregation and analysis of team-level performance metrics such as win/loss ratios, average goals scored/conceded per game.

– **Performance Metrics:** Evaluation of advanced metrics like Expected Goals (xG), Expected Goals Against (xGA), and possession percentages.

Club Analysis: This section delves into the economic and competitive aspects of clubs. It assesses market value fluctuations of clubs and players, comparing club performances across different leagues and competitions to understand their standing in various contexts.

– **Market Value Assessment:** Analysis of the market value of players and clubs over time, including factors influencing market fluctuations.

– **Comparison of Club Performances:** Comparative analysis of club performances across different leagues, seasons, and competitions.

Game Event Analysis: This analysis targets the in-game dynamics, such as scoring patterns, player attributes influence on performance, and the overall impact of key players on match outcomes.

- **Scoring Patterns:** Identification of temporal patterns in scoring, such as periods during a match when a team is more likely to score.
- **Correlations Between Player Attributes and Performance Metrics:** Analysis of how specific player attributes (e.g., speed, stamina) correlate with performance outcomes.
- **Impact Evaluation:** Assessment of the impact of key players on game outcomes and overall team performance.

Transfer Market Analysis: Player transfers are scrutinized to uncover market trends and assess how the movement of players between clubs influences team performance in future seasons.

- **Player Transfer Analysis:** Examination of player transfer data to identify trends and patterns in the transfer market.
- **Impact of Transfers on Team Performance:** Analysis of how player transfers influence team performance in subsequent seasons.

Formation Analysis: A study of how different tactical formations influence game outcomes, identifying trends and their effects on key performance indicators such as goals scored or conceded.

- **Team Performance by Formations:** Analysis of team performance under different tactical formations (e.g., 4-4-2, 3-5-2).
- **Trends in Formations:** Identification of formation trends across clubs and leagues over time.
- **Impact on Key Performance Indicators (KPIs):** Evaluation of how different formations impact key performance metrics like goals scored or conceded.

Comparison of International Games with League Games: This comparison examines the differences in player and team performance between international and domestic league matches, focusing on contextual factors that may influence the outcomes.

- **Performance Metrics Comparison:** Comparative analysis of player and team performance in international matches versus domestic league matches.
- **Contextual Differences:** Exploration of contextual differences that may influence performance in international versus league games.

Player Tracking: This analysis uses real-time positional data and detailed statistics to examine player movements and contributions on the pitch, as well as historical performance trends across seasons.

- **Real-time Positional Data:** Analysis of player positions on the pitch during a game to understand movement patterns and positional play.
- **In-depth Performance Metrics:** Detailed statistics such as passes completed, tackles made, shots on goal, and distance covered to provide a comprehensive view of player contributions.
- **Detailed Game Events:** Analysis of specific game events, including goal times, substitutions, fouls, assists, and injuries to understand their impact on the game's outcome.
- **Historical Performance:** Examination of player performance statistics over multiple seasons to identify trends and career progression.

Detailed Team Tactics and Formations: The focus here is on the tactical adjustments teams make during a match, particularly formation changes and strategic plays, to respond to different game situations.

- **Formation Changes During the Game:** Analysis of how and when teams change formations during a match to adapt to different situations.
- **Strategic Plays:** Evaluation of strategic plays executed during the game, such as set-pieces and tactical shifts.

Injury and Suspension Data: This category tracks injury and suspension data, assessing their effects on individual and team performances, as well as player fitness levels.

- **Player Injuries:** Documentation and analysis of injury data to understand common injury types and their impact on player and team performance.

- **Suspensions:** Analysis of suspension data to assess its effect on team dynamics and match outcomes.
- **Fitness Levels:** Evaluation of player fitness levels and their correlation with performance metrics.

Fan and Social Media Sentiment: Analysis of fan engagement and public sentiment on social media platforms to understand how external factors like public opinion may influence team morale and performance.

- **Fan Engagement:** Analysis of fan engagement data from social media platforms to gauge public sentiment and its potential influence on player and team morale.
- **Public Sentiment:** Evaluation of public sentiment surrounding players, teams, and matches using sentiment analysis tools.

Training and Practice Data: Training data is examined to understand how practice routines impact player development and match performance, providing insights into areas needing improvement.

- **Training Sessions:** Analysis of training session data to understand practice routines and their effectiveness in improving performance.
- **Practice Match Performance:** Evaluation of performance in practice matches to identify potential areas of improvement.
- **Player Development Metrics:** Documentation of player development metrics to track progress and development over time.

Revenue and Financial Performance: This category analyzes the financial aspects of football, including club revenues, financial health, and trends in transfer transactions to assess the financial sustainability of clubs over time.

- **Club Revenues:** Analysis of club revenues from various sources such as ticket sales, merchandise, and sponsorships.
- **Financial Health:** Evaluation of the overall financial health of clubs, including profitability and financial sustainability.
- **Transfer Transactions:** Detailed analysis of individual transfer transactions, including dates, fees involved, and the financial impact on clubs.
- **Financial Trends:** Identification of trends in financial transactions over time, such as increasing transfer fees and spending patterns.

4.3 Common Data Model

The common data model is a relational database schema that has been merged from a number of different datasets. The dataset **Football Data from Transfermarkt** serves as the basis for this model. Subsequently, additional attributes and tables were incorporated. The final result of the data model is illustrated in Figure 1. A detailed graphical representation of the model is also provided in Appendix A.

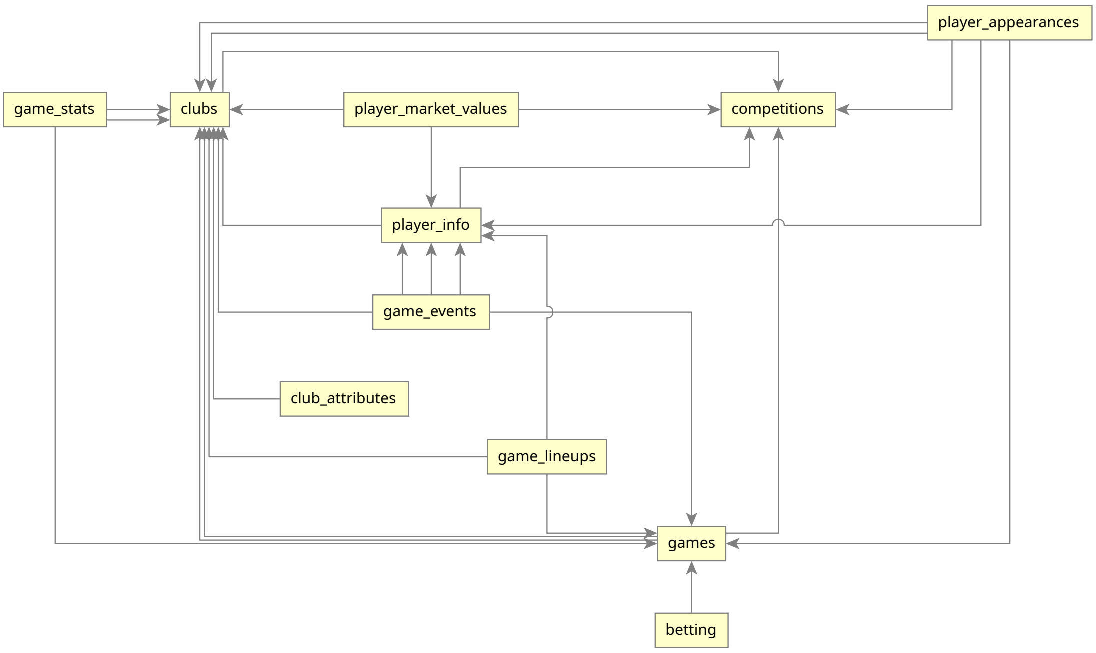

Figure 1: Schema, displaying table names only.

The Appendix B provides a complete overview of the tables included in each relation. The merged data model presented integrates information from the open football data sources. The data model is systematically organized into several key section, taking an high impact from the **Football Data From Transfermarkt** dataset:

- **Player Appearances**: This section captures data on individual player performances in matches. It includes fields such as appearance identifiers, game and player IDs, club affiliations, match dates, player names, and performance metrics like goals, assists, minutes played, and disciplinary records.
- **Game Statistics**: Focusing on match outcomes and team performances, this section comprises fields related to game identifiers, club IDs, goals scored

by both the team and opponents, match results, and managerial information. It provides essential insights into team dynamics and match outcomes.

- **Clubs**: Detailed information about football clubs is consolidated here, including club identifiers, names, codes, domestic competition affiliations, financial valuations, squad characteristics, and managerial details. This section aids in analyzing club-level strategies and financial health.

- **Competitions**: This part encompasses data on various football competitions, covering competition identifiers, names, types, associated countries, and relevant dates. It situates matches and performances within the broader context of specific tournaments and leagues.

- **Game Events**: Recording significant occurrences during matches, this section includes data on events like goals, substitutions, and cards, along with their timing and involved players or teams. It allows for granular analysis of match progressions and pivotal moments.

- **Game Lineups**: Information on player participation in matches is detailed here, including starting lineups, substitutes, positions, and roles such as team captains. This section supports studies on tactical formations and player utilization.

- **Games**: This comprehensive section aggregates match-level information, including game identifiers, competition affiliations, seasons, rounds, dates, participating teams, scores, venues, and officiating details. It serves as a foundational dataset for match analysis.

- **Player Market Values**: Focusing on the economic aspects, this section records players' market valuations over time, providing insights into transfer market dynamics and player valuation trends.

- **Player Information**: Detailed player profiles are presented here, covering personal information, physical attributes, playing positions, and skill assessments. This data is crucial for performance analysis and talent scouting.

- **Betting Data**: Incorporating betting odds and related information, this section enables the examination of betting market behaviors and the relationship between predicted and actual match outcomes.

- **Club Attributes**: This part includes tactical and strategic attributes of clubs, such as playing styles and formation tendencies, contributing to analyses of team strategies and their effectiveness.

The integration of these datasets enriches the data model by combining the strengths and unique contributions of each source. For instance, **Football Data From Transfermarkt** provides extensive coverage across multiple sections, offering detailed player and club information. The **European Soccer Database** adds depth with detailed player skill ratings and betting data, while other sources like **WoSo Stats** and **Worldcup** contribute specific details on stadium capacities and international competitions, respectively.

4.4 Mapping Datasets to Analytical Tasks

A map was constructed to assess the availability of data required for various football analysis tasks within the developed common data model. This visualization in Appendix C maps each analysis task to the necessary data fields and indicates their presence in the data model. The results reveal areas where the data model sufficiently supports analysis activities and highlight significant gaps where essential data is missing. The map categorizes the analysis tasks into thirteen main groups, each with specific subtasks. The availability of the required data fields is indicated using a color-coded system as well as written text:

- **Available**: All required data fields are present.
- **Partially Available**: Some required data fields are present.
- **Not Available**: Required data fields are absent.

4.5 Feasibility Evaluation

The feasibility evaluation of the open football datasets revealed mixed results in terms of their utility for various analytical tasks. Each dataset was assessed using several criteria, including data size, frequency of updates, documentation, usability of the data structure, and temporal coverage.

4.5.1 Data Size

The datasets examined, such as the European Soccer Database, Football (Soccer) Data for Everyone, and Transfermarkt Datasets, vary significantly in size. Larger datasets provide comprehensive player and match data spanning multiple seasons, leagues, and competitions. These datasets are well-suited for statistical analysis and machine learning applications due to their scale, offering millions of data points for detailed examination. However, the larger the dataset, the more computational resources are required, which may present challenges in terms of storage and processing for some users. Smaller datasets, such as OpenLigaDB, provide valuable but more limited data, focusing on specific leagues or time periods. While useful for niche analyses, these smaller datasets may not offer the breadth required for more generalized tasks, such as global player market analysis or multi-league performance comparisons.

4.5.2 Frequency of Updates

The frequency of dataset updates varies widely across the sources. Automated and API-based datasets tend to be updated frequently, often on a match-by-match basis. On the other hand, manually curated datasets, such as those found on International football results from 1872 to 2024, are often updated less frequently, sometimes only annually or biannually. For real-time analysis, such as in-game prediction models or live tactical evaluations, datasets with frequent updates are crucial. However, the inconsistency in update frequency across datasets means that merging data from multiple sources can be problematic. Automated processes that rely on real-time data may struggle to integrate datasets that are updated less frequently, creating potential gaps in analysis.

4.5.3 Documentation

The availability and quality of documentation also vary among the datasets. Publicly accessible datasets, like the International football results from 1872 to 2024 and Transfermarkt Datasets, typically include extensive documentation, detailing variables, collection methods, and preprocessing steps. Well-documented datasets ensure that researchers can confidently utilize the data, understanding its limitations and context. However, some datasets, such as European Soccer Database or Football (Soccer) Data for Everyone, lack comprehensive documentation, which complicates their use in more advanced analytics. The absence of clear explanations regarding variable definitions or data collection methodologies poses risks, especially in tasks requiring high accuracy, such as injury prediction or transfer market evaluations.

4.5.4 Usability of Data Structure

The usability of the data structure is critical for effective analysis, particularly when integrating multiple datasets. The structure of datasets like Transfermarkt Datasets and WorldCup is well-defined, with clear relationships between tables and consistent data types. These datasets are conducive to integration with existing analysis frameworks, making them ideal for complex tasks, such as player tracking or tactical analysis. However, the structure of some datasets requires manual intervention to standardize formats and resolve inconsistencies. For instance, WoSo Stats may use unique and inconsistent formats, which complicates the process of merging them with larger datasets. Automated merging of such datasets is currently not feasible without significant manual effort, as data structures often differ in terms of key variables, metadata, and formats.

4.5.5 Time Range

The time range covered by each dataset significantly impacts its usability for longitudinal studies. Datasets such as the International Football Results Dataset provide extensive historical data, covering more than a century of match results, which is invaluable for trend analysis and the development of long-term predictive models. Conversely, some datasets focus only on recent seasons or specific competitions, limiting their use in broader analyses that require historical context. For instance, OpenLigaDB provides more recent data for specific leagues, making it less useful for tasks that require comprehensive historical coverage, such as the analysis of long-term player development or club performance trends.

4.6 Automation of Dataset Integration

Automating the integration of multiple datasets remains a significant challenge due to the varying formats, structures, and update frequencies across sources. While some automation tools can facilitate merging by standardizing data formats, the frequent changes and inconsistencies in datasets - such as alterations to schema, naming conventions, or update intervals - often require manual intervention. For instance, integrating datasets like Transfermarkt Datasets with other sources, such as Football (Soccer) Data for Everyone or football.db, involves aligning differing variable definitions and addressing gaps in data coverage. These inconsistencies introduce significant complexity into the automation process, making fully automated integration unreliable.

5 Discussion

In this discussion, we synthesize the findings of our research to address the central and sub-research questions outlined earlier. We evaluate the current state of openly accessible football datasets, examining their types, attributes, and suitability for various analytical tasks in football data science. By exploring the typical tasks performed using these datasets, identifying gaps between available data and analytical needs, and assessing the potential for automation in data collection, we provide a comprehensive understanding of both the strengths and limitations of open football datasets. This analysis not only answers each research question but also highlights areas for improvement and future research in the field.

In addressing the main research question - *What is the current state of open football datasets and their utilizations?* - the research indicates that while there is a substantial amount of openly accessible football data available, it primarily supports basic analytical tasks. The datasets are suitable for foundational analyses such as player performance tracking and basic team metrics. However, significant limitations exist in the availability of advanced metrics and comprehensive data required for more in-depth analyses. The utilization of these datasets is often constrained by gaps in data, inconsistencies in quality, and legal restrictions related to data collection methods like web scraping. Consequently, the current state of open football datasets is one of partial utility; they are helpful for basic analyses but inadequate for advanced, comprehensive studies in football data science.

5.1 RQ1: Types and Attributes of Openly Accessible Football Datasets

The research identified a variety of openly accessible football datasets available on the internet, each with distinct types and attributes. These datasets primarily include player statistics, match results, team formations, and basic performance metrics. Notable examples are the **Transfermarkt Datasets** and the **European Soccer Database**, which offer detailed and reliable data for top European leagues. These datasets encompass information such as goals, assists, minutes played, and disciplinary records. The attributes of these datasets vary in quality and completeness. While some provide comprehensive and high-quality data suitable for foundational analyses, others exhibit inconsistencies in metadata, lack details about data sources, and have incomplete or outdated information. The majority of open datasets are sourced through web scraping, which often leads to issues with data quality, licensing restrictions, and legal considerations.

5.2 RQ2: Typical Tasks Performed on Football Datasets

The typical tasks performed on football datasets involve player performance tracking, team formations analysis, match outcome predictions, and basic team performance metrics analysis such as goals scored and conceded. Analysts and researchers utilize these datasets to study tactical formations, player contributions, and to make comparative analyses between different leagues or competitions. Formation analysis allows for investigations into team tactics and their influence on match outcomes. Additionally, some datasets enable basic predictive modeling for match results based on historical performance data.

5.3 RQ3: Gaps between Open Datasets and Tasks

The results indicate that the current data model effectively supports several key areas of football analysis. Tasks that are fully supported include player performance metrics such as games played, minutes, goals, assists, and disciplinary records. Basic team performance statistics, including goals scored and conceded, as well as match results, are also well-supported. Additionally, formation analysis allows for studies on team tactics and their influence on performance, while comparative analysis between international and league games is possible based on available performance metrics. However, certain tasks are only partially supported. For instance, club market value assessments lack data on influencing factors, and correlation studies between player attributes and performance metrics are limited by incomplete attribute data. Detailed game events contain some data but are not fully comprehensive, and player development metrics over time are insufficient for in-depth analysis. Several tasks are entirely unsupported. Advanced performance metrics such as possession percentages are absent, which restricts deeper analytical insights. Transfer market analysis is not feasible due to the lack of transfer data, and tactical analyses involving formation changes during games and strategic plays cannot be conducted. Additionally, data on injuries, suspensions, and player fitness are missing, limiting studies on player welfare and availability. Fan engagement and public sentiment analysis are not possible as there is no social media data, and training or practice data is unavailable, which prevents the analysis of training effectiveness and player development. Lastly, financial performance analysis is unfeasible due to the absence of revenue and financial health data.

5.4 RQ4: Automation in Collecting Open Football Datasets

The extent to which automation can be employed in collecting open football datasets is limited by several challenges. While web scraping techniques allow for the automated collection of data from various online sources, inconsistencies in data formats, structures, and licensing restrictions pose significant obstacles. Integrating disparate data sources into a unified framework is complicated by the lack of standardization and completeness in metadata. Moreover, the semi-open or gray datasets often lack transparency regarding update schedules and licensing conditions, making it difficult to automate data collection processes reliably. Therefore, while automation is possible to some extent, it requires careful consideration of legal, ethical, and technical factors to ensure the data collected is both usable and compliant with relevant regulations.

6 Conclusion and Recommendations

Through a comprehensive survey of open football datasets and the analysis of their utilizations, we have identified several key findings which we will briefly explain in the following:

Availability and Accessibility of Open Football Datasets. A variety of open football datasets exist, including well-known ones like *football.db*, *Transfermarkt Datasets*, and the *European Soccer Database*. These datasets primarily cover top European leagues, international competitions, and specific cup tournaments. They offer essential data for player performance, match results, and team statistics. However, while some datasets are openly accessible under permissive licenses, others are semi-open or restricted, which affects their usability for broader research.

Types of Tasks Supported. The datasets support foundational football analysis tasks such as player performance tracking (goals, assists, and minutes played), team performance metrics, and basic formation analysis. Tasks such as scoring patterns, market value assessments, and tactical analyses of formations can be conducted with the available data.

Gaps in Data Coverage. Despite the availability of various datasets, critical gaps exist. Notably, advanced performance metrics such as possession percentages, detailed game events, and player-specific attributes are either missing or incomplete in many datasets. The lack of this data restricts more in-depth analyses, such as correlating player attributes with performance, detailed injury prediction models, and studies involving transfer market dynamics. Furthermore, social media sentiment analysis and training data are absent, which limits research on fan engagement and its impact on team morale, as well as player development through training sessions.

Licensing and Usage Challenges. The diversity of licensing conditions poses challenges for the use and modification of datasets. While some datasets are open under licenses like CC0 or MIT, others impose restrictions on redistribution or commercial use, which hampers the integration of different data sources into a unified framework for more comprehensive research.

Automation of Dataset Collection. Automating the collection of open football datasets, though feasible, faces challenges such as disparate formats, incomplete metadata, and inconsistent update frequencies across different sources. This complicates the process of building a cohesive, comprehensive dataset for longitudinal studies.

Mapping Datasets to Analytical Tasks. The common data model developed as part of this study highlights the potential for integrating multiple datasets to support a wider range of analytical tasks. However, the mapping process revealed that some advanced tasks remain unsupported due to incomplete or missing data.

These findings reveal that while current open football datasets provide a good foundation for basic football analysis, significant gaps remain in terms of data completeness, consistency, and scope, limiting the potential for more advanced research.

6.1 Summary of Contributions to the Field

This thesis contributes to the field of football analytics and open data research in several key areas. *First*, it provides a comprehensive survey of open football datasets, evaluating their accessibility, quality, and usability for football analysis. This survey serves as a valuable resource for researchers aiming to utilize publicly available data in their work. *Second*, the development of a common data model for integrating various football datasets introduces a standardized method for organizing and analyzing football data, making it easier to align datasets with common analytical tasks and facilitating more advanced analyses. *Third*, the research also identifies critical gaps in current datasets, particularly in advanced metrics, tactical analysis, and data licensing, offering clear directions for future improvements. *Finally*, the thesis explores the feasibility of automating dataset collection, addressing the technical challenges associated with maintaining up-to-date football datasets, which is crucial for advancing real-time football analytics.

6.2 Recommendations for Researchers, Practitioners, and Data Providers to Address Identified Gaps

This thesis proposes several recommendations for researchers, practitioners, and data providers to address the identified gaps in open football datasets. It is crucial for researchers to recognise the limitations of existing datasets and implement strategies to address these gaps. In instances where more sophisticated metrics, such as possession percentages or tactical shifts, are necessary, researchers should explore the use of supplementary data sources or consider developing new methods of data collection. It is essential to collaborate with data providers and practitioners to guarantee that datasets remain current and meet the evolving analytical requirements, particularly in areas such as player development, fitness, and market dynamics.

For practitioners, including football clubs and analysts, advocating for comprehensive data-sharing practices that incorporate advanced metrics –such as player movements, tactical formations, and player wellness data– can significantly enhance the quality of analysis and strategy development. It would also be beneficial for practitioners to consider investing in their own data collection technologies, such as GPS tracking and real-time video analysis, in order to bridge the gaps identified in publicly available datasets. For data providers, standardising data formats and improving the availability of metadata can enhance the integration of datasets and their overall utility for research purposes, thereby increasing the value of the data they provide. It is also essential to clarify and, where possible, relax licensing restrictions in order to encourage broader use and innovation. It is recommended that open licenses, such as CC0 or MIT, be used to provide researchers with the flexibility to utilise, modify and integrate data. Finally, data providers should extend the reach of their datasets to include underrepresented areas such as lower-tier leagues, youth competitions, and non-European tournaments, thereby facilitating more diverse and inclusive football research.

A Appendix

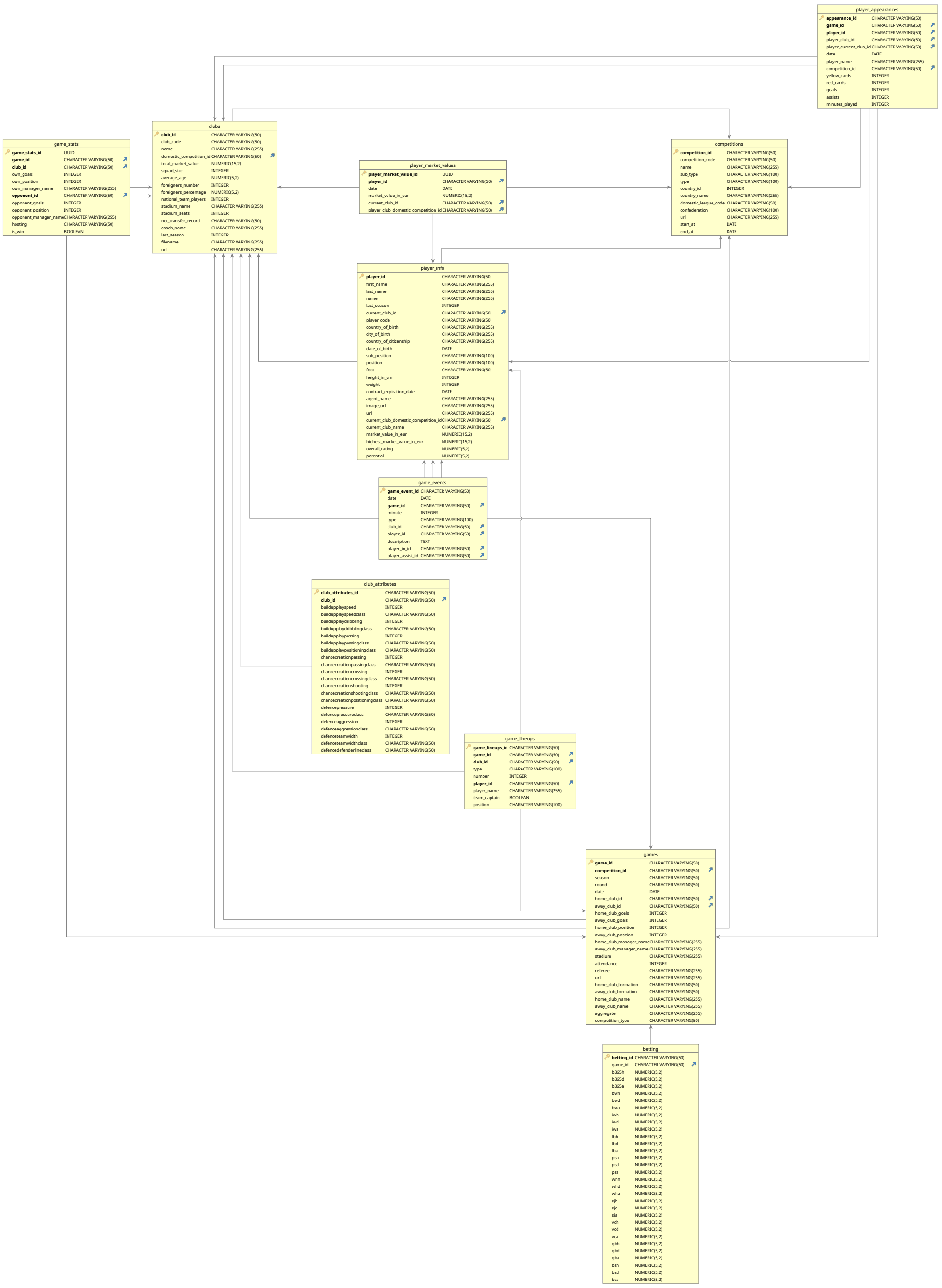
player_appearances
appearance_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)
player_id CHARACTER VARYING(50)
player_club_id CHARACTER VARYING(50)
player_current_club_id CHARACTER VARYING(50)
date DATE
player_name CHARACTER VARYING(255)
competition_id CHARACTER VARYING(50)
yellow_cards INTEGER
red_cards INTEGER
goals INTEGER
assists INTEGER
minutes_played INTEGER
game_stats
game_stats_id UUID
game_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
own_goals INTEGER
own_position INTEGER
own_manager_name CHARACTER VARYING(255)
opponent_id CHARACTER VARYING(50)
opponent_goals INTEGER
opponent_position INTEGER
opponent_manager_name CHARACTER VARYING(255)
hosting CHARACTER VARYING(50)
is_win BOOLEAN
clubs
club_id CHARACTER VARYING(50)
club_code CHARACTER VARYING(50)
name CHARACTER VARYING(255)
domestic_competition_id CHARACTER VARYING(50)
total_market_value NUMERIC(15,2)
squad_size INTEGER
average_age NUMERIC(5,2)
foreigners_number INTEGER
foreigners_percentage NUMERIC(5,2)
national_team_players INTEGER
stadium_name CHARACTER VARYING(255)
stadium_seats INTEGER
net_transfer_record CHARACTER VARYING(50)
coach_name CHARACTER VARYING(255)
last_season INTEGER
filename CHARACTER VARYING(255)
url CHARACTER VARYING(255)
player_market_values
player_market_value_id UUID
player_id CHARACTER VARYING(50)
date DATE
market_value_in_eur NUMERIC(15,2)
current_club_id CHARACTER VARYING(50)
player_club_domestic_competition_id CHARACTER VARYING(50)
competitions
competition_id CHARACTER VARYING(50)
competition_code CHARACTER VARYING(50)
name CHARACTER VARYING(255)
sub_type CHARACTER VARYING(100)
type CHARACTER VARYING(100)
country_id INTEGER
country_name CHARACTER VARYING(255)
domestic_league_code CHARACTER VARYING(50)
confederation CHARACTER VARYING(100)
url CHARACTER VARYING(255)
start_at DATE
end_at DATE
player_info
player_id CHARACTER VARYING(50)
first_name CHARACTER VARYING(255)
last_name CHARACTER VARYING(255)
name CHARACTER VARYING(255)
last_season INTEGER
current_club_id CHARACTER VARYING(50)
player_code CHARACTER VARYING(50)
country_of_birth CHARACTER VARYING(255)
city_of_birth CHARACTER VARYING(255)
country_of_citizenship CHARACTER VARYING(255)
date_of_birth DATE
sub_position CHARACTER VARYING(100)
position CHARACTER VARYING(100)
foot CHARACTER VARYING(50)
height_in_cm INTEGER
weight INTEGER
contract_expiration_date DATE
agent_name CHARACTER VARYING(255)
image_url CHARACTER VARYING(255)
url CHARACTER VARYING(255)
current_club_domestic_competition_id CHARACTER VARYING(50)
current_club_name CHARACTER VARYING(255)
market_value_in_eur NUMERIC(15,2)
highest_market_value_in_eur NUMERIC(15,2)
overall_rating NUMERIC(5,2)
potential NUMERIC(5,2)
game_events
game_event_id CHARACTER VARYING(50)
date DATE
game_id CHARACTER VARYING(50)
minute INTEGER
type CHARACTER VARYING(100)
club_id CHARACTER VARYING(50)
player_id CHARACTER VARYING(50)
description TEXT
player_in_id CHARACTER VARYING(50)
player_assist_id CHARACTER VARYING(50)
club_attributes
club_attributes_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
buildupplayspeed INTEGER
buildupplayspeedclass CHARACTER VARYING(50)
buildupplaydribbling INTEGER
buildupplaydribblingclass CHARACTER VARYING(50)
buildupplaypassing INTEGER
buildupplaypassingclass CHARACTER VARYING(50)
buildupplaypositioningclass CHARACTER VARYING(50)
chancecreationpassing INTEGER
chancecreationpassingclass CHARACTER VARYING(50)
chancecreationcrossing INTEGER
chancecreationcrossingclass CHARACTER VARYING(50)
chancecreationshooting INTEGER
chancecreationshootingclass CHARACTER VARYING(50)
chancecreationpositioningclass CHARACTER VARYING(50)
defencepressure INTEGER
defencepressureclass CHARACTER VARYING(50)
defenceaggression INTEGER
defenceaggressionclass CHARACTER VARYING(50)
defenceteamwidth INTEGER
defenceteamwidthclass CHARACTER VARYING(50)
defencedefenderlineclass CHARACTER VARYING(50)
game_lineups
game_lineups_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
type CHARACTER VARYING(100)
number INTEGER
player_id CHARACTER VARYING(50)
player_name CHARACTER VARYING(255)
team_captain BOOLEAN
position CHARACTER VARYING(100)
games
game_id CHARACTER VARYING(50)
competition_id CHARACTER VARYING(50)
season CHARACTER VARYING(50)
round CHARACTER VARYING(50)
date DATE
home_club_id CHARACTER VARYING(50)
away_club_id CHARACTER VARYING(50)
home_club_goals INTEGER
away_club_goals INTEGER
home_club_position INTEGER
away_club_position INTEGER
home_club_manager_name CHARACTER VARYING(255)
away_club_manager_name CHARACTER VARYING(255)
stadium CHARACTER VARYING(255)
attendance INTEGER
referee CHARACTER VARYING(255)
url CHARACTER VARYING(255)
home_club_formation CHARACTER VARYING(50)
away_club_formation CHARACTER VARYING(50)
home_club_name CHARACTER VARYING(255)
away_club_name CHARACTER VARYING(255)
aggregate CHARACTER VARYING(255)
competition_type CHARACTER VARYING(50)
betting
betting_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)
b365h NUMERIC(5,2)
b365d NUMERIC(5,2)
b365a NUMERIC(5,2)
bwh NUMERIC(5,2)
bwd NUMERIC(5,2)
bwa NUMERIC(5,2)
iwh NUMERIC(5,2)
iwd NUMERIC(5,2)
iwa NUMERIC(5,2)
lbh NUMERIC(5,2)
lbd NUMERIC(5,2)
lba NUMERIC(5,2)
psh NUMERIC(5,2)
psd NUMERIC(5,2)
psa NUMERIC(5,2)
whh NUMERIC(5,2)
whd NUMERIC(5,2)
wha NUMERIC(5,2)
sjh NUMERIC(5,2)
sjd NUMERIC(5,2)
sja NUMERIC(5,2)
vch NUMERIC(5,2)
vcd NUMERIC(5,2)
vca NUMERIC(5,2)
gbh NUMERIC(5,2)
gbd NUMERIC(5,2)
gba NUMERIC(5,2)
bsh NUMERIC(5,2)
bsd NUMERIC(5,2)
bsa NUMERIC(5,2)

Figure 2: Schema, displaying all attributes.

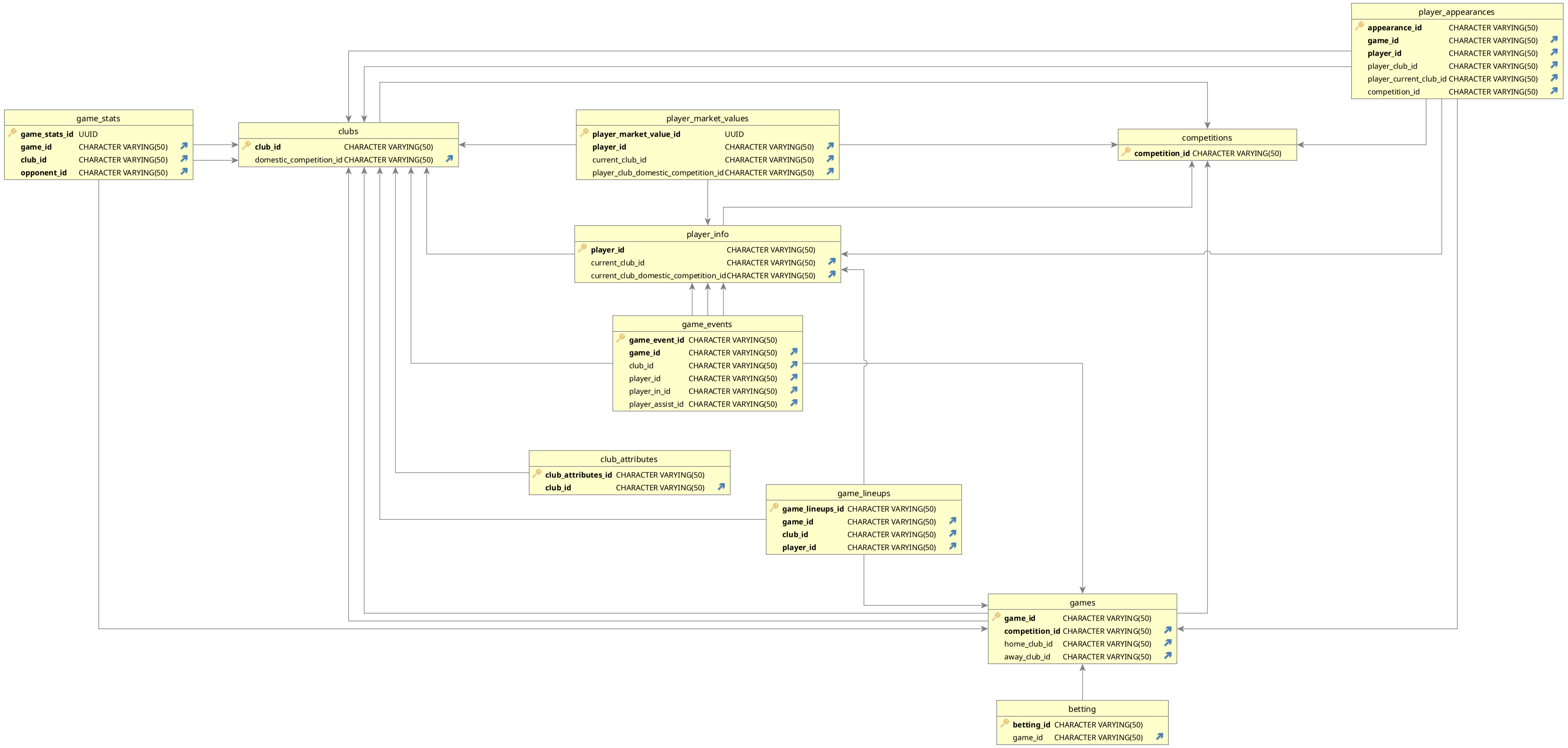
player_appearances
appearance_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)
player_id CHARACTER VARYING(50)
player_club_id CHARACTER VARYING(50)
player_current_club_id CHARACTER VARYING(50)
competition_id CHARACTER VARYING(50)
game_stats
game_stats_id UUID
game_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
opponent_id CHARACTER VARYING(50)
clubs
club_id CHARACTER VARYING(50)
domestic_competition_id CHARACTER VARYING(50)
player_market_values
player_market_value_id UUID
player_id CHARACTER VARYING(50)
current_club_id CHARACTER VARYING(50)
player_club_domestic_competition_id CHARACTER VARYING(50)
competitions
competition_id CHARACTER VARYING(50)
player_info
player_id CHARACTER VARYING(50)
current_club_id CHARACTER VARYING(50)
current_club_domestic_competition_id CHARACTER VARYING(50)
game_events
game_event_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
player_id CHARACTER VARYING(50)
player_in_id CHARACTER VARYING(50)
player_assist_id CHARACTER VARYING(50)
club_attributes
club_attributes_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
game_lineups
game_lineups_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)
club_id CHARACTER VARYING(50)
player_id CHARACTER VARYING(50)
games
game_id CHARACTER VARYING(50)
competition_id CHARACTER VARYING(50)
home_club_id CHARACTER VARYING(50)
away_club_id CHARACTER VARYING(50)
betting
betting_id CHARACTER VARYING(50)
game_id CHARACTER VARYING(50)

Figure 3: Schema, displaying primary key and foreign key only.

B Appendix

Field	TM	ESDB	FDB	IFR	ODB	WSS	WCD
Player Appearances							
appearance_id	x						
game_id	x						
player_id	x						
player_club_id	x						
player_current_club_id	x						
date	x	x				x	
player_name	x			x		x	
competition_id	x						
yellow_cards	x			x			
red_cards	x			x			
goals	x			x			
assists	x						
minutes_played	x						
Game Stats							
game_id	x						
club_id	x						
own_goals	x	x	x	x	x	x	x
own_position	x						
own_manager_name	x						
opponent_id	x						
opponent_goals	x	x	x	x	x	x	x
opponent_position	x						
opponent_manager_name	x						
hosting						x	
is_win	x	x	x	x	x	x	x
Clubs							
club_id	x						
club_code	x	x	x				x
name	x	x			x		x
domestic_competition_id	x						
total_market_value	x						
squad_size	x						
average_age	x						
foreigners_number	x						

Continued on next page

Table 9 – continued from previous page

Field	TM	ESDB	FDB	IFR	ODB	WSS	WCD
foreigners_percentage	x						
national_team_players	x						
stadium_name	x			x		x	
stadium_seats						x	
net_transfer_record	x						
coach_name	x						
last_season	x						
filename	x						
url		x		x		x	x
Competitions							
competition_id	x						x
competition_code	x						
name	x	x	x	x	x		x
sub_type	x						
type	x			x			
country_id	x	x					
country_name	x	x					
domestic_league_code	x						
confederation	x						
url	x						
start_at			x				
end_at			x				
Game Events							
game_event_id	x						
date	x			x			x
game_id	x						
minute	x	x	x	x		x	x
type	x	x	x	x		x	x
club_id	x						
player_id	x						
description	x				x		x
player_in_id	x						
player_assist_id	x						
Game Lineups							
game_lineups_id	x						
game_id	x						

Continued on next page

Table 9 – continued from previous page

Field	TM	ESDB	FDB	IFR	ODB	WSS	WCD
club_id	x						
type	x						
number	x					x	
player_id	x		x				
player_name	x					x	
team_captain	x						
position	x					x	
Games							
game_id	x						
competition_id	x						
season	x	x					x
round	x		x			x	x
date	x		x			x	x
home_club_id	x						
away_club_id	x						
home_club_goals	x	x	x	x	x	x	x
away_club_goals	x	x	x	x	x	x	x
home_club_position	x						
away_club_position	x						
home_club_manager_name	x					x	x
away_club_manager_name	x					x	x
stadium	x						x
attendance	x						x
referee	x						x
url	x						
home_club_formation	x						
away_club_formation	x						
home_club_name	x						
away_club_name	x						
aggregate	x	x	x	x	x	x	x
competition_type	x						
Player Market Values							
player_id	x						
date	x						
market_value_in_eur	x						
current_club_id	x						

Continued on next page

Table 9 – continued from previous page

Field	TM	ESDB	FDB	IFR	ODB	WSS	WCD
player_club_domestic_competition_id	x						
Player Info							
player_id	x						
first_name	x		x				x
last_name	x		x				x
name	x		x				x
last_season	x						
current_club_id	x						
player_code	x						
country_of_birth	x						x
city_of_birth	x						x
country_of_citizenship	x						
date_of_birth	x	x	x				
sub_position	x						
position	x		x				x
foot	x						
height_in_cm	x	x					
weight		x					
contract_expiration_date	x						
agent_name	x						
image_url	x						
url	x						x
current_club_domestic_competition_id	x						
current_club_name	x		x				
market_value_in_eur	x						
highest_market_value_in_eur	x						
overall_rating		x					
potential		x					
preferred_foot		x					
attacking_work_rate		x					
defensive_work_rate		x					
crossing		x					
finishing		x					
heading_accuracy		x					
short_passing		x					
volleys		x					

Continued on next page

Table 9 – continued from previous page

Field	TM	ESDB	FDB	IFR	ODB	WSS	WCD
dribbling		x					
curve		x					
free_kick_accuracy		x					
long_passing		x					
ball_control		x					
acceleration		x					
sprint_speed		x					
agility		x					
reactions		x					
balance		x					
shot_power		x					
jumping		x					
stamina		x					
strength		x					
long_shots		x					
aggression		x					
interceptions		x					
positioning		x					
vision		x					
penalties		x					
marking		x					
standing_tackle		x					
sliding_tackle		x					
gk_diving		x					
gk_handling		x					
gk_kicking		x					
gk_positioning		x					
gk_reflexes		x					
Betting							
betting_id							
game_id							
B365H		x					
B365D		x					
B365A		x					
BWH		x					
BWD		x					

Continued on next page

Table 9 – continued from previous page

Field	TM	ESDB	FDB	IFR	ODB	WSS	WCD
BWA		x					
IWH		x					
IWD		x					
IWA		x					
LBH		x					
LBD		x					
LBA		x					
PSH		x					
PSD		x					
PSA		x					
WHH		x					
WHD		x					
WHA		x					
SJH		x					
SJD		x					
SJA		x					
VCH		x					
VCD		x					
VCA		x					
GBH		x					
GBD		x					
GBA		x					
BSH		x					
BSD		x					
BSA		x					
Club Attributes							
club_attributes_id							
club_id							
buildUpPlaySpeed		x					
buildUpPlaySpeedClass		x					
buildUpPlaySpeedClass		x					
buildUpPlayDribbling		x					
buildUpPlayDribblingClass		x					
buildUpPlayPassing		x					
buildUpPlayPassingClass		x					
buildUpPlayPositioningClass		x					

Continued on next page

Table 9 – continued from previous page

Field	**TM**	**ESDB**	**FDB**	**IFR**	**ODB**	**WSS**	**WCD**
chanceCreationPassing		x					
chanceCreationPassingClass		x					
chanceCreationCrossing		x					
chanceCreationCrossingClass		x					
chanceCreationShooting		x					
chanceCreationShootingClass		x					
chanceCreationPositioningClass		x					
defencePressure		x					
defencePressureClass		x					
defenceAggression		x					
defenceAggressionClass		x					
defenceTeamWidth		x					
defenceTeamWidthClass		x					
defenceDefenderLineClass		x					

C Appendix

Table 10: Data Availability for Analysis Tasks

Analysis Task	**Required Data Fields**	**Availability**
1. Player Performance Analysis		
a. Games/Minutes Played	Player ID, Game ID, Minutes Played, Performance Metrics	Available
b. Scores and Assists	Player ID, Game ID, Goals, Assists, Date	Available
c. Disciplinary Records	Player ID, Game ID, Yellow Cards, Red Cards	Available
2. Team Performance Analysis		
a. Calculation of Team Statistics	Game ID, Club ID, Goals Scored/Conceded, Match Result	Available
b. Performance Metrics (Advanced)	Expected Goals (xG), xGA, Possession Percentages	Available
3. Club Analysis		
a. Market Value Assessment	Player/Club Market Values Over Time, Influencing Factors	Partially Available
b. Comparison of Club Performances	Club Performance Metrics, Competition IDs, Season Data	Available
4. Game Event Analysis		
a. Scoring Patterns	Game Events with Timestamps, Event Type	Available

Continued on next page

Analysis Task	Required Data Fields	Availability
b. Correlations Between Player Attributes and Performance Metrics	Player Attributes, Performance Metrics	Partially Available
c. Impact Evaluation	Player Performance Data, Game Results	Available
5. Transfer Market Analysis		
a. Player Transfer Analysis	Transfer Dates, Fees, Clubs Involved	Not Available
b. Impact of Transfers on Team Performance	Transfer Data, Team Performance Metrics	Not Available
6. Formation Analysis		
a. Team Performance by Formations	Team Formations, Match Results	Available
b. Trends in Formations	Team Formations Over Time	Available
c. Impact on KPIs	Formations, KPIs (goals scored/conceded)	Available
7. Comparison of International Games with League Games		
a. Performance Metrics Comparison	Game Data (Intl vs. Domestic), Performance Metrics	Available
b. Contextual Differences	Contextual Factors (travel distance, rest days)	Not Available
8. Player Tracking		
a. Real-time Positional Data	Real-time GPS Data	Not Available
b. In-depth Performance Metrics	Passes Completed, Tackles, Shots on Goal, Distance Covered	Not Available
c. Detailed Game Events	Specific Game Events (fouls, injuries)	Partially Available

Continued on next page

Analysis Task	**Required Data Fields**	**Availability**
d. Historical Performance	Player Performance Over Seasons	Available
9. Detailed Team Tactics and Formations		
a. Formation Changes During the Game	Formation Changes with Timestamps	Not Available
b. Strategic Plays	Data on Set-Pieces, Tactical Shifts	Not Available
10. Injury and Suspension Data		
a. Player Injuries	Injury Incidents, Types, Recovery Time	Not Available
b. Suspensions	Suspension Details	Not Available
c. Fitness Levels	Fitness Metrics	Not Available
11. Fan and Social Media Sentiment		
a. Fan Engagement	Social Media Metrics, Sentiment Data	Not Available
b. Public Sentiment	Sentiment Scores	Not Available
12. Training and Practice Data		
a. Training Sessions	Training Session Details	Not Available
b. Practice Match Performance	Practice Match Data	Not Available
c. Player Development Metrics	Development Over Time	Partially Available
13. Revenue and Financial Performance		
a. Club Revenues	Revenue Sources (ticket sales, merchandise)	Not Available

Continued on next page

Analysis Task	**Required Data Fields**	**Availability**
b. Financial Health	Profitability, Financial Sustainability	Not Available
c. Transfer Transactions	Transfer Dates, Fees Involved	Not Available
d. Financial Trends	Historical Financial Data	Not Available

D Appendix

- **Player Performance Analysis**
  - Analysis of individual player stats such as games/minutes played, scores, assists, cards, etc.
  - **Paper**:
    - *A public data set of spatio-temporal match events in soccer competitions* Pappalardo et al. (2019)
  - **Platform**:
    - `https://onefootball.com/de/home`
    - `https://www.squawka.com/en/`
    - `https://www.whoscored.com/`
- **Team Performance Analysis**
  - Calculation of team statistics, including win/loss records, goals, and team rankings.
  - **Papers**:
    - *Exploring and Modelling Team Performances of the Kaggle European Soccer Database* Carpita et al. (2019)
    - *The Open International Soccer Database for Machine Learning* Bergmann et al. (2013)
  - **Platform**:
    - `https://www.whoscored.com/`
    - `https://www.soccerstats.com/`
- **Club Analysis**
  - Examination of club performance, market value, and comparison between clubs.
  - **Market Value**:
    - **Platform**: `https://www.transfermarkt.com/`
  - **Club Performance Comparison**:
    - *Exploring and Modelling Team Performances of the Kaggle European Soccer Database* Carpita et al. (2019)
    - **Platform**: `https://www.squawka.com/en/`

- **Game Event Analysis**
  - Detailed analysis of game events, such as when a team scores and player contributions.
  - **Papers**:
    - *Exploring and Modelling Team Performances of the Kaggle European Soccer Database* Carpita et al. (2019)
  - **Tasks**:
    - Identify correlations between player attributes and performance metrics
    - Evaluate the impact of individual players on game outcomes
- **Transfer Market Analysis**
  - Examining player transfers, identifying trends, and assessing the impact on team performance.
  - **Platform**:
    - `https://www.transfermarkt.com/`
  - **Paper**:
    - *Does the Stock Market Take into Consideration Football Players' Injuries?* Mrhari and Hasssouni (2023a)
- **Formation Analysis**
  - Analysis of team formations and their impact on performance.
  - **Tasks**:
    - Performance based on formations
    - Formation trends across clubs and leagues
    - Impact of formations on key performance indicators (KPIs) like goals scored or conceded
  - **Platform**:
    - `https://www.whoscored.com/`
- **Comparison of International Games vs League Games**
  - Evaluating the differences between international and league performances.
  - **Paper**:

        - *Big Data and Tactical Analysis in Elite Soccer: Future Challenges and Opportunities for Sports Science* Watanabe et al. (2021)

    - **Platform**:
        - `https://www.soccerstats.com/`

- **Player Tracking**
    - Tracking player movements, game events, and using video-based data to improve analytics.
    - **Papers**:
        - *Automatic Event Detection in Football Using Tracking Data* Huang et al. (2022)
        - *Data-Driven Visual Performance Analysis in Soccer: An Exploratory Prototype* Benito Santos et al. (2018)
        - *Open Dataset Recorded by Single Cameras for Multi-Player Tracking in Soccer Scenarios* Huang et al. (2022)
        - *Setting a Baseline for Long-Shot Real-Time Player and Ball Detection in Soccer Videos* Moutselos and Maglogiannis (2023)

References

Bai, Z. and Bai, X. (2021). Sports big data: Management, analysis, applications, and challenges. *Complexity*, 2021:1–11.

Baumer, B. and Zimbalist, A. (2014). The sabermetric revolution: Assessing the growth of analytics in baseball. *Economics: Faculty Books.*

Benito Santos, A., Therón, R., Losada, A., Sampaio, J., and Peñas, C. (2018). Data-driven visual performance analysis in soccer: An exploratory prototype. *Frontiers in Psychology*, 9:2416.

Bergmann, T., Bunk, S., Eschrig, J., Hentschel, C., Knuth, M., Sack, H., and Schüler, R. (2013). Generating a linked soccer dataset. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 146–149, New York, NY, USA. Association for Computing Machinery.

Burton, N., Bradish, C., and Dempsey, M. (2019). Exploring expatriate fan identification in international football supporters. *Sport, Business and Management: An International Journal*, 9(1):78–96.

Carpita, M., Ciavolino, E., and Pasca, P. (2019). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, 19.

Chandra, B., Shinny, J., Adhitya, K. M., et al. (2024). Prediction of football player performance using machine learning algorithm. *PREPRINT (Version 1) available at Research Square.*

Ćwiklinski, B., Gielczyk, A., and ChoraŚ, M. (2021). Who will score? a machine learning approach to supporting football team building and transfers. *Entropy*, 23.

Herberger, T. A. and Litke, C. (2021). The impact of big data and sports analytics on professional football: A systematic literature review. In Herberger, T. A. and Dötsch, J. J., editors, *Digitalization, digital transformation and sustainability in the global economy*, pages 147–171. Springer International Publishing.

Huang, W., He, S., Sun, Y., Evans, J., Song, X., Geng, T., Sun, G., and Fu, X. (2022). Open dataset recorded by single cameras for multi-player tracking in soccer scenarios. *Applied Sciences.*

Liu, H., Hopkins, W., and Ruano, M. (2015). Modelling relationships between match events and match outcome in elite football. *European Journal of Sport Science*, pages 1–10.

Moutselos, K. and Maglogiannis, I. (2023). Setting a baseline for long-shot real-time player and ball detection in soccer videos. *ArXiv*, abs/2311.06892.

Mrhari, E. and Hasssouni, M. (2023a). Does stock market take into consideration football playersâ€™ injuries? *Research Papers in Economics and Finance*, 7:89–100.

Mrhari, E. M. and Hasssouni, M. (2023b). Does stock market take into consideration football players' injuries? *Research Papers in Economics and Finance*, 7(1):89–100.

Owusu, G. (2008). Ai and computer-based methods in performance evaluation of sporting feats: an overview. *Artificial Intelligence Review*, 27:57–70.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., and Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6.

Payyappalli, V. M. and Zhuang, J. (2019). A data-driven integer programming model for soccer clubs' decision making on player transfers. *Environment Systems and Decisions*, 39(4):466–481.

Prieto-Lage, I., Argibay-González, J. C., Paramés-González, A., Pichel-Represas, A., Bermúdez-Fernández, D., and Gutiérrez-Santiago, A. (2021). Patterns of injury in the spanish football league players. *International Journal of Environmental Research and Public Health*, 19(1):252.

Watanabe, N. M., Shapiro, S., and Drayer, J. (2021). Big data and analytics in sport management. *Journal of Sport Management*, 35(3):197 – 202.