UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Educação Física

MURILO MERLIN

NEW APPROACH TO ANALYZE THE PASSING IN SOCCER MATCHES USING MULTIVARIATE AND MACHINE LEARNING TECHNIQUES

NOVA ABORDAGEM PARA ANÁLISE DO PASSE EM JOGOS DE FUTEBOL USANDO TÉCNICAS MULTIVARIADAS E APRENDIZAGEM DE MÁQUINA

CAMPINAS
2020

MURILO MERLIN

NEW APPROACH TO ANALYZE THE PASSING IN SOCCER MATCHES USING MULTIVARIATE AND MACHINE LEARNING TECHNIQUES

NOVA ABORDAGEM PARA ANÁLISE DO PASSE EM JOGOS DE FUTEBOL USANDO TÉCNICAS MULTIVARIADAS E APRENDIZAGEM DE MÁQUINA

Thesis presented to the Faculty of Physical Education, University of Campinas in partial fulfillment of the requirements for the degree of Doctor, in the area of Biodynamics of Movement and Sport.

Tese apresentada à Faculdade de Educação Física da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Educação Física, na Área de Biodinâmica do Movimento e Esporte.

Orientador: SERGIO AUGUSTO CUNHA

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO MURILO MERLIN, E ORIENTADA PELO PROF. DR. SERGIO AUGUSTO CUNHA.

CAMPINAS
2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Educação Física
Dulce Inês Leocádio - CRB 8/4991

M548n

Merlin, Murilo, 1980-
New approach to analyse the passing in soccer matches using multivariate and machine learning techniques / Murilo Merlin. – Campinas, SP : [s.n.], 2020.

Orientador: Sergio Augusto Cunha.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Educação Física.

1. Passe (Futebol). 2. Futebol. 3. Rastreamento (posição). 4. Análise multivariada. 5. Inteligência artificial. I. Cunha, Sergio Augusto. II. Universidade Estadual de Campinas. Faculdade de Educação Física. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Nova abordagem para análise do passe em jogos de futebol usando técnicas multivariadas e aprendizagem de máquina
Palavras-chave em inglês:
Passing (Soccer)
Soccer
Tracking (position)
Multivariate analysis
Artificial intelligence
Área de concentração: Biodinâmica do Movimento e Esporte
Titulação: Doutor em Educação Física
Banca examinadora:
Sergio Augusto Cunha [Orientador]
Allan da Silva Pinto
Paulo Regis Caron Ruffino
Felipe Arruda Moura
Paulo Roberto Pereira Santiago
Data de defesa: 19-11-2020
Programa de Pós-Graduação: Educação Física

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-6156-498
- Currículo Lattes do autor: http://lattes.cnpq.br/6628862094423521

COMISSÃO EXAMINADORA[1]

Prof. Dr. Sergio Augusto Cunha
Orientador

Prof. Dr. Allan da Silva Pinto
Membro Titular da Banca

Prof. Dr. Felipe Arruda Moura
Membro Titular da Banca

Prof. Dr. Paulo Regis Caron Ruffino
Membro Titular da Banca

Prof. Dr. Paulo Roberto Pereira Santiago
Membro Titular da Banca

[1]Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

DEDICATÓRIA

Ao meu saudoso e amado pai Domingos Batista Merlin por dedicar sua vida para proporcionar aos seus filhos o que não teve oportunidade de fazer, estudar... não deixando, no entanto, de ter uma mente brilhante.

AGRADECIMENTOS

Em primeiro lugar agradeço ao meu amigo, professor e orientador Sergio Cunha pela oportunidade e suporte no desenvolvimento dessa pesquisa. Uma vez dentro do programa de doutorado, convivi com pessoas que durante a caminhada de alguma forma, seja conceitual, metodológica, instrumental ou emocional contribuíram para a construção e desenvolvimento dessa ideia. Na Unicamp, agradeço especialmente aos amigos e colaboradores do LIB e LAES, especialmente Alexandre, Kleber e Monezi, aos encontros às terças com o Prof. Barreto, à parceria valiosíssima com o Instituto de Computação e as orientações do Ricardo Torres e Allan, entre outros. Agradeço ao suporte do Laboratório de Biomecânica da UEL através do Felipe Moura e sua equipe, assim como o Laboratório de Biomecânica e Controle Motor da USP – Ribeirão Preto através do Paulo Santiago (Preto) e equipe. Agradeço ao Figueirense Futebol Clube, em especial ao Fisiologista Tiago Cetolin que abriu as portas do clube para troca de conhecimentos e nossa coleta de dados.

Durante o programa de doutorado tive a oportunidade de desenvolver parte da pesquisa no Laboratório de *Performance Analysis in Team Sports* na Universidade Trás-os-Montes e Alto Douro em Portugal sob orientação do Jaime Sampaio. Durante esse período tive um salto significativo em meu aprendizado e no desenvolvimento de parte da pesquisa. Agradeço, portanto, a todos os colaboradores do laboratório em especial ao Jaime pela oportunidade, Bruno Gonçalves, Juliana e todos os demais amigos e colaboradores. Foram tempos necessários e determinantes em minha caminhada.

Nada seria possível se "fora das quatro linhas" não houvessem pessoas importantes que me dessem suporte em um dos períodos emocionalmente mais desafiadores da minha vida. Agradeço pelo convívio dos amigos em Campinas, em especial as acolhidas na casa da família Cunha. E por fim, agradeço a base de qualquer pessoa, a família. Agradeço minha Mãe e irmãos por respeitar minha decisão não tão lógica ao deixar meu trabalho para mergulhar nessa missão, e ao UNIVERSO pela oportunidade em resgatar aquilo que deixei lá atrás e que hoje ressignificam minha vida, minha família, Gabi e Zeca.

Essa pesquisa foi apoiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), Código 001" e pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) sob o Grants [#2016/50250-1, #2017/20945-0, #2018/19007-9, #2019/16253-1, #2019/17729-0, e #2019/22262-3].

ABSTRACT

Traditionally, in the practical and scientific context, the variables used to explain individual and collective performance related to the ability to perform passes provide insufficient information to improve performance. The main variables used as a performance indicator when analyzing the passes are the percentage of successful passes and time of ball possession in the case of pass sequences. The general objective of this research was to propose a new approach to analyze the passing in soccer matches using multivariate and machine learning techniques. The specific objectives were proposed based on two types of analysis: ball possession (BP) analysis (Study 1), and passing analysis (Studies 2 and 3). Were analyzed four official matches of the Brazilian Soccer Championship 2016. In study 1 were used 41 'notational', ‘space occupation’, and 'displacement synchronization’ predictor variables. The BP were classified into three groups using clustering techniques: short, medium and long. Fisher discriminant analysis (FDA) identified the five most relevant variables to describe each group, suggesting collective behaviors that help to maintain BP and perform passes. The second and third study focused their analyzes on the concept of passing difficulty, that guided the proposition of 36 variables related to the pass, pressure on the passing player, pressure on the passing receiver, ball trajectory, pitch position and passing player techniques. In both studies, we used a sample with 465 passes labelled by experts. The passes were classified such low difficulty (Low-DP), medium difficult (Medium-DP) and high difficulty (High-DP). In study 2, the FDA presented 72.0% of accuracy when classifying the degree of passing difficulty into three classes. In addition, were identified 16 between 32 variables that best explain the degree of passing difficulty in soccer. In study 3, we improved the prediction by classifying passes using machine learning algorithms. The support vector machine (SVM), a non-linear model, reaching a balanced accuracy of 88% in their best performance. Then we use this model to predict a new sample of passes (total = 2,522), and to analyze players e positions. The High-DP had a success rate of 49.3% only, followed by 84.0% for Medium-DP and 94.9% for Low-DP. These variables were used as input in the principal component analysis (PCA). The principal component (PC1) showed a higher correlation with the variable accuracy in High-DP and Medium-DP, suggesting that it is more important to consider the player's ability to complete High-PD and Medium-DP than Low-DP. In addition, the PC1 scores were used to rank the best passing players. The models and variables propose can be used by coaches in a practical context to analyze passing performance of their players and teams, in order to improve performance in performing passes, considering that the passes are the important determinants of success in soccer matches.

Keywords: passing (soccer); soccer; tracking (position); multivariate analysis; artificial intelligence.

RESUMO

Tradicionalmente, no contexto prático e científico, as varáveis usadas para explicar desempenho individual e coletivo relacionados a capacidade de executar passes fornecem informações insuficientes para melhora do desempenho. As principais variáveis usadas como indicador de desempenho ao analisar passes são percentuais de passes bem sucedidos e tempo de posse de bola, no caso de sequência de passes. O objetivo geral desta pesquisa foi propor uma nova abordagem para analisar o passe em partidas de futebol utilizando técnicas multivariadas e de aprendizagem de máquina (ML). Os objetivos específicos foram propostos com base em dois tipos de análise: análise da posse de bola (BP) (Estudo 1), e análise do passe (Estudos 2 e 3). Foram analisadas quatro partidas oficiais do Campeonato Brasileiro de Futebol 2016. No estudo 1 foram usadas 41 variáveis preditoras 'notacionais', 'ocupação do espaço' e 'sincronização de deslocamento'. As BP foram classificadas em três grupos por meio de técnicas de agrupamento: curta, média e longa. A análise discriminante de Fisher (FDA) identificou cinco variáveis mais relevantes para descrever cada grupo, sugerindo comportamentos coletivos que ajudam a manter a BP e realizar passes. O segundo e o terceiro estudos focaram suas análises no conceito de dificuldade de passe, que norteou a proposição de 36 variáveis relacionadas ao passe, pressão sobre o passador, pressão sobre receptor do passe, trajetória da bola, posição no campo e técnicas do passador. Em ambos os estudos, usamos uma amostra com 465 passes rotulados por experts. Os passes foram classificados como baixa dificuldade (Low-DP), média dificuldade (Medium-DP) e alta-dificuldade (High-DP). No estudo 2, a FDA apresentou 72,0 % de acurácia ao classificar o grau de dificuldade de passagem em três classes. Além disso, foram identificadas 16 entre 32 variáveis que melhor explicam o grau de dificuldade do passe no futebol. No estudo 3, melhoramos a previsão classificando os passes usando algoritmos de ML. O *support vector machine* (SVM), modelo não linear, alcançou acurácia balanceada de 88% em seu melhor desempenho. Em seguida, usamos esse modelo para predizer uma nova amostra de passes (total = 2.522) e analisar jogadores e posições. High-DP apresentou apenas 49,3% de passes bem sucedidos, seguido por 84,0% para Medium-PD e 94,9% para Low-PD. Essas variáveis foram usadas como inputs na análise de componentes principais (PCA). A principal componente (PC1) apresentou maior correlação com a variável acertos em High-DP e Medium-DP, sugerindo que é mais importante considerar a capacidade do jogador em executar passes de alta e média dificuldade do que passes de baixa dificuldade. Além disso, os scores da PC1 foram usados para classificar os melhores passadores. Os modelos e variáveis propostos podem ser usados por treinadores em um contexto prático para analisar seus jogadores e equipes, a fim de melhorar o desempenho na execução de passes, visto que os passes são determinantes do sucesso em partidas de futebol.

Palavras-chave: passe (futebol); futebol; rastreamento (posição); análise multivariada, inteligência artificial.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. General introduction

1.1. Match Analysis in soccer

The process of match analysis in soccer (Association Football) is a long time and has been improved over time influenced by technological, computational, sports science and data science evolution. In general, the analyses applied in soccer matches have focused on descriptive (activity patterns of players), comparative (playing position competitive level, contextual variables) and predictive (probability to score a goal, probability of the game result) analyses, being that predictive analysis represents a very small proportion (Sarnento et al., 2014).

Match analysis supplies information through the interaction of the various factors involved: technical, tactical, mental and psychological factors, which provide an understanding of the situations that lead to success in sport (Carling, Reilly, & Williams, 2009). Contemporary match analysis systems provide a rich source of quantitative data that allows the identification of key performance indicators (biomechanical, technical, tactical or behavioral) individual or for a team, being for a match or season (Carling et al., 2009). Performance indicators are a selection of action variables, single or combined, that try to define the aspects of a performance (Hughes & Bartlett, 2010). The information obtained from the data favors the planning and direction of the training process in order to improve individual and collective performance, besides to contribute to the evolution of soccer (Figure 1.1). Sports performance analysis enables the coach, players and the managers to objectively assess and thereby improve their sporting performance (Kumar, 2014).

We can highlight at least four phases with specific tendencies in the match analysis process in soccer: 1) Tendencies in notational analyzes with a focus on technical actions; 2) Tendencies in the analysis of physical demand using spatiotemporal data; 3) Tendencies in the tactical analysis using spatiotemporal data and proposition of new metrics; 4) Tendencies in predictive analysis using multivariate and machine learning techniques from large volumes and variety of data.

In one of the first and most classic studies on game analysis in football, (Reep & Benajmin, 1968) analyzed 3,213 English league matches between the years 1953 and 1968 using notational technique and related the occurrence of goals with different length

of the sequence of passes. In the first phase, historically the match analysis in soccer has focused on technical demand based on notational techniques and inferences using frequency, density, order and accuracy of actions.

From the 2,000 years, started the development and/or adaptation of technological resources aimed at obtaining spatiotemporal data, such a multiple camera semi-automatic systems, local position measurement (LPM) technology and global positioning system (GPS) technology (Buchheit, 2014; Carling et al., 2008). The position and time data of each player allows obtaining various information related to the different aspects of the match, physical, technical and tactical, adding physiological variables that constitute an important set of interdependent variables (Rein & Memmert, 2016).

From the access to spatiotemporal data, the second phase of the process, the tendencies in sports research was the interest in understanding the physical demand based on activity profiles of soccer players such distances covered, speed thresholds, accelerations, relationships between internal and external loads, among others (Baron et al., 2006; Barros et al., 2007; Bradley et al., 2011, 2013; Clemente et al., 2013; Mclaren et al., 2017; Osgnach et al., 2009).

Subsequently, in the third phase of the match analysis process in soccer, increase the interest in tactical aspects. Spatiotemporal data can be used to develop collective performance indicators capable of describing and understanding the dynamics of the match, for example: measuring inter-player coordination, measuring inter-team coordination before critical events, and measuring team-team interaction and compactness coefficients (Memmert, Lemmink, & Sampaio, 2016). Some metrics have been proposed such coverage area and spread on the pitch (Moura, Martins, Anido, Barros, & Cunha, 2012), centroid (Sampaio & Maças, 2012), control and space creation models (Couceiro, Martins, Figueiredo, Mendes, & Clemente, 2014; Fernandez & Bornn, 2018; Fujimura & Sugihara, 2005), synchronization (Folgado, Duarte, Fernandes, & Sampaio, 2014; Folgado, Duarte, Marques, Gonçalves, & Sampaio, 2018).

New challenges and opportunities have emerged from FIFA's permission to use electronic resources in official matches, and with the growth of the market for technologies for match analysis. The research tendencies aimed at analyzing the soccer matches started to explore more contextualized problems, integrating different aspects of the match as tactical, technical, physical and physiological. For this purpose, predictive analyzes using multivariate and machine learning techniques from large volumes and variety data have gained more space. Rein & Memmert (2016) discussed this topic in the

article entitled “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science”. The authors consider tactics as a central component of elite soccer, therefore, to represent the complex processes underlying the team's tactical behavior requires detailed data on technical aspects, physiological performance, teams' position system, among others. This new perspective, contextual and multivariate, has required joint efforts from different areas, such as sports scientists and data scientists (Goes et al., 2020). For example, approaches using techniques from deep reinforcement learning to valuate multiplayer positionings based on positional data (Dick & Brefeld, 2019), predicting match outcome using tactical performance metrics computed from position tracking data (Goes et al., 2019), among others.

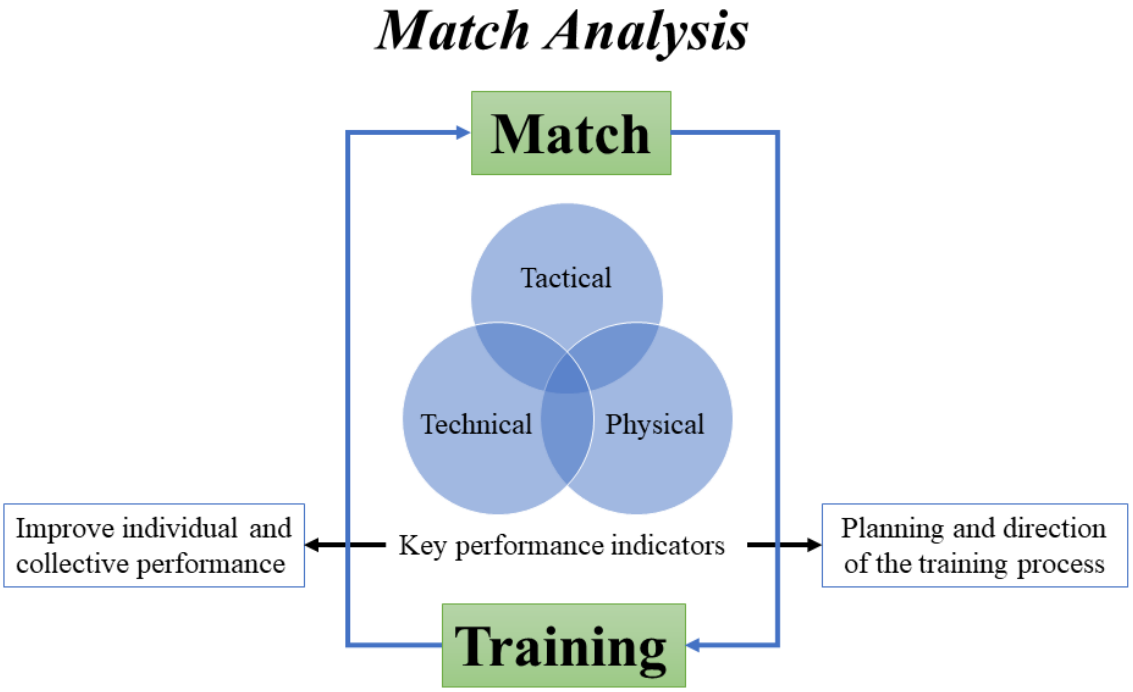


Figure 1.1. Diagram illustrating the main objective of the match analysis process in soccer.

1.2. Multivariate and Machine Learning techniques

Learning sciences has a fundamental role in the statistics, data mining (DM) and artificial intelligence (AI) fields, intersecting with engineering and other disciplines. Multivariate statistics comprises a set of methods used in situations in witch several variables are measured simultaneously in each sample element, the variables are correlated, in general, and the complexity increases as the number of variables increases (Mining, 2008). The

propose of machine learning methods is the same as that of statistical ones. Both aim at improving forecasting accuracy by minimizing some loss function, typically the sum of squared errors (Makridakis, Spiliotis, & Assimakopoulos, 2018). The aim of Machine Learning (ML) is to study, engineer, and improve mathematical models which can be trained with context-related data, to infer the future and to make decisions without complete knowledge of all influencing elements, i.e., an agent adopts a statistical learning approach, trying to determine the right probability distributions and use them to compute the action (value or decision) that is most likely to be successful (with the least error) (Bonaccorso, 2017).

In general, the structuring of the data set in both cases, statistics and ML is similar, where *n* quantitative and/or qualitative independent variables (inputs) are assigned to each sample element, which are associated with their dependent variable (outputs), that can be quantitative (regression models) or qualitative (classification models). In ML, independent variables are more commonly called "predictor variables" or "features", and the dependent variable is the "response variable". The objective is to use the inputs from a sample set, to predict the outputs values or classification (Mining, 2008).

Prediction models can be based on linear or non-linear techniques. Normally ML methods use non-linear algorithms, while statistical methods use linear processes (Makridakis et al., 2018).

Recently, the increase of sports-related data available in terms of volume, velocity, and variety of data, the big data characterization (Riahi & Riahi, 2018), has required joint efforts from different areas, such as sports scientists and data scientists (Goes et al., 2020; Rein & Memmert, 2016). As a consequence, the application of multivariate and machine learning (ML) techniques has increased considerably, with important contributions in performance analysis, such injury prevention (Sikka, Baer, Raja, Stuart, & Tompkins, 2019), strategy analysis (Dick & Brefeld, 2019), training design and talent identification (Goes et al., 2020), game style and game system identification (Dick & Brefeld, 2019; Fernandez-navarro et al., 2016), identifying indicators of success (Kite & Nevill, 2017; Whitaker, Silva, & Edwards, 2018), even prediction of technical actions such as passes, which will be discussed in the next topic.

1.3. Passing analysis in soccer matches

From the historical perspective described and the current scenario, i.e., four phases with specific tendencies in the match analysis process in soccer previously highlighted, the pass has always been one of the most studied elements of the match, since the first study cited about match analysis, until to more recent studies.

In our view, the pass is the basis of soccer matches. Soccer matches have become more complex, faster, and players frequently need to work on reduced space to maintain ball possession (Wallace & Norton, 2014). This action, therefore, has been the main resource used to comply with the match offensive principles, i.e., maintain possession, progress in the pitch and create space and opportunity for scoring (Ouellette, 2004). In addition, it has been considered one of the key performance indicators (Cintia et al., 2015; Goes et al., 2019, 2018). The pass is the most used action by the player with ball possession, representing 69% of the ball actions (Bransen & Haaren, 2019). On average, a typical match comprises 500 passes per team (Goes et al., 2018). This means that approximately every 10s on average, a player has the control of the ball to execute the pass. The relevance of this action as a determinant for performance within the match is unquestionable.

The pass has been analyzed from two perspectives. The first perspective, as a passes sequence or ball possession. A passing sequence for one team possession was defined in terms of sequence length (Hughes & Franks, 2005). Ball possession was deemed to start when a player on the analyzed team had sufficient control over the ball to enable a deliberate influence on its direction (Jones, 2004). In the present study, we analyze ball possession from the perspective of a passes sequence, considering that passes are the most used action when the team has control of the ball. The second perspective, as a discrete event, i.e., with a technical action in which it is usually classified as successful or unsuccessful. The pass in soccer was defined as the deliberate act of touching and projecting the ball on de pitch to another teammate control over it, maintaining the possession of the team (Cunha, Moura, Santiago, Castellani, & Barbieri, 2011; Wallace & Norton, 2014; Horton, Gudmundsson, Chawla, & Estephan, 2014).

Traditionally, in the practical and research context, the variables used to explain individual and collective performance related to the ability to perform passes provide insufficient information to improve performance. For example, in the ball possession case, the main variable used to assess ball possession as a performance indicator is the

time of possession. The most research insist in an attempt to establish a cause-effect relationship, ie, how ball possession's time influences performance indicators such as shots and goals or performance across the season (Collet, 2013; Hughes & Franks, 2005), or considering aspects such as passing frequency, pitch zones where the ball moves, passing characteristics and match status (Cintia, Giannotti et al., 2015; Lago & Martín, 2007; Jones, 2004; Paixão et al., 2015a). In the case of passes analysis, the player and team performance normally are determined by the accuracy, i.e., percentage of successful passes, which supposedly indicates the player and team efficiency. In addition,
passes analysis in soccer matches has focused on inferences using frequency, density, and order of actions (Chassy, 2013; Gyarmati et al., 2014; Hughes & Franks, 2005; Lago & Martín, 2007; Mitschke & Milani, 2014; Peña & Navarro, 2015; Reep & Benajmin, 1968).

Currently, with access to spatiotemporal data, the pass has been analyzed in a more contextualized, multivariate and predictive perspective. Accurate information for all players of both teams, favors proposals for new metrics (Goes et al., 2018; Gyarmati & Stanojevic, 2016; Rein, Raabe, & Memmert, 2017), index (Cintia et al., 2015), network analysis (Gonçalves et al., 2017; Mchale & Relton, 2018) and predictive analysis (Bransen & Haaren, 2019; Horton, Gudmundsson, Chawla, & Estephan, 2014; I. Mchale, 2015; Power, Ruiz, Wei, & Lucey, 2017; Spearman, Basye, Dick, Hotovy, & Pop, 2017). Similarly, there is researches with ball possession using contextualized and multivariate approach, how the proposed of a new metric "dangerousity" to quantify the probability of goal at each moment of ball possession (Link, Lang, & Seidenschwarz, 2016) and collective and regularity variables to discriminate between short and long BP (Aguiar, Gonçalves, Botelho, Duarte, & Sampaio, 2017).

Improving the relevance of pass information in soccer matches is undoubtedly a promising path, especially in a multivariate perspective, based on spatiotemporal data. Although there has been considerable progress, still needed new proposals about the question, considering the importance of the pass into the match context.

Based on these assumptions, from the access to the spatiotemporal data, and with the multivariate and machine learning techniques available, this research amid to break with analyzes traditionally used in the scientific and practical context and propose a new approach for analyzing passes in soccer matches. We intend to collaborate with specific problem based on two perspectives described about passes analysis: the passes as a sequence or ball possession and the pass as a single event.

In relation of the passes into the ball possession (BP), in our view, more important that to relate BP's properties to performance indicators, is identify and describe collective behaviors that help to maintain BP and perform passes, considering their relevance to the match. Identify and describe collective behaviors that help to maintain BPS and perform passes is more promising than insist in an attempt to establish a cause-effect relationship.

When analyzing the pass as an event, we definitely need to rupture from analyzes based only on accuracy, successful or unsuccessful passes. The degree of difficulty of the action has been overlooked in the literature. The information of the degree of difficulty of each pass in the match would allow coaches to analyze the efficiency by relativizing the action difficulty. In our view, the pass is a technical-tactical action in which the difficulty depends on the interaction of several technical factors (e.g., body position and orientation, ball contact, movement speed, and pass distance) and tactical (e.g., team interaction and space occupation by individual players, group, or by the team), to the ball reaches its destination. Considering that the passing difficulty has a multivariate nature, it would be important to identify and discuss the variables that best explain this phenomenon. In addition, the classification of passes in different degrees of difficulty could enable us to discriminate players, position, and offensive sequences, taking into consideration the merit of successfully executing highly complex actions.

In summary, the present thesis sought to contribute with two specific problems related to the passes analysis in soccer matches, considering as premises: the passes as the basis in soccer matches because are the most frequent action and are an important performance indicator; improve the relevance of the information on the passes based on the previous premise; propose contextualized analyzes based on spatiotemporal data and using multivariate and machine learning techniques available.

1.4. Overview

This thesis was organized based on three original articles published and/or submitted to international journals. The Ethics Committee of the Campinas State University approved this research, protocol CAAE 56582616.8.0000.5404 (appendix A). Original images of the data collection are shown in Appendix B and comparison between real (TV camera) and 2D image in Appendix C. The general objective of the thesis was to propose a new approach to analyze the passing in soccer matches using multivariate and machine learning techniques. The three studies that compose this document are related to the

specific objectives of the thesis. The first study, Session I, *Exploring the determinants of success in different clusters of ball possession sequences in soccer*, published article (Appendix D), was developed in partnership with the University Trás-os-Montes e Alto Douro in Portugal during exchange financed by Santander Program (VRERI n° 065/2017). This study analyzed the pass within ball possession (BP) sequences using 41 variables predominantly collective and dynamic. The main objective was identify which tactical variables most discriminate the different BP. The second and third study, Session II and III respectively, focused their analyses on concept of passing difficulty, originally proposed in this thesis: "passing difficulty refers to the degree of technical and tactical demands that the passing player has to complete the action successfully". This concept guided the proposition of 36 variables related to the pass, pressure on the passing player, pressure on the passing receiver, ball trajectory, pitch position and passing player techniques. The study two, *Classification and determinants of passing difficulty in soccer: a multivariate approach*, aimed to classify automatically the degree of passing difficulty in soccer matches e identify and discuss the variables that most explain the passing difficulty using spatiotemporal data. The study three, *Who are the best passing players in professional soccer? Machine learning approach classifies passes with different levels of difficulty and discriminate the best passing players*, aimed to improve the prediction by classifying passes using machine learning algorithms, and to apply the model with the best performance to discriminate players and positions. In summary, the thesis is structured with the following textual elements: General Introduction, Objective, Session I, Session II, Session III, General Discussion and General Conclusion.

2. Objective

2.1.General objective

The general objective of this research was to propose a new approach to analyze the passing in soccer matches using multivariate and machine learning techniques.

2.2. Specific objectives

The specific objectives were proposed based on two types of analysis. The first with a focus on passing sequence analysis (Study 1), and the second in analyzing the pass as an event (Studies 2 and 3):

Study 1: i) classify ball possession sequences according to the duration and number of passes; ii) identify which tactical variables most discriminate the different ball possession sequences, as classified in the previous step.

Study 2: (i) classify automatically the degree of passing difficulty in soccer matches; (ii) identify and discuss the variables that most explain the passing difficulty using spatiotemporal data.

Study 3: (i) to classify automatically the degree of passing difficulty in soccer matches using machine learning classifiers; (ii) to apply the model with the best performance to discriminate players, positions, and offensive sequences.

Session I

3. Session I (Study 1)

Exploring the determinants of success in different clusters of ball possession sequences in soccer

ABSTRACT

The purpose of this study was two-step: classify ball possession (BP) according to the duration and number of passes; identify which tactical variables most discriminate the different BP. We obtained 527 BPs from four official matches of the Brazilian Soccer Championship 2016. Forty-one 'notational', ‘space occupation’, and 'displacement synchronization’ predictor variables were used. The BPs were classified into three groups: short (11.07 ± 4.49s, 1.93 ± 0.99 passes), medium (26.83 ± 7.33s, 5.41 ± 1.84 passes), long (55.50 ± 14.97s, 12.11 ± 4.61 passes). Discriminant analysis identified the five most relevant variables to describe each group: coefficient of variation (CV) of the defensive team’s synchronization-Y, CV defensive team´s synchronization-X, successful pass last third, CV distance between offensive team’s centroid and target, mean of the offensive team’s width. The approach highlights important variables and could benefit the description of offensive and defensive game sequences to provide precise knowledge on the process.

Keywords: ball possession; tactical; multivariate; soccer

3.1. Introduction

Ball possession (BP) is the consequence of interactions determined by contextual factors, such as quality of opponent, tactical configuration, match status, or venue of the match (Link, Hoernig, Nassis, Laughlin, & Witt, 2017). Tactically, controlling the ball possession as much possible consists of a substantial set of on-ball and off-ball actions to generate scoring chances. Some of these actions are associated with game principles like creating numerical superiority or promoting disorder on the opponents' defense, but most importantly, generating and occupying spaces (Fernandez & Bornn, 2018).

Although BP is a complex phenomenon whose success depends on the combination of many variables, most research insist in an attempt to establish a cause-effect relationship, ie, how BP's time influences performance indicators such as shots and goals or performance across the season (Collet, 2013; M. Hughes & Franks, 2005). Besides that, literature studies have explored others properties of BP, considering aspects such as passing frequency, pitch zones where the ball moves, passing characteristics and match status (Cintia, Giannotti et al., 2015; Lago & Martín, 2007; P. D. Jones, 2004; Paixão et al., 2015a).

In our viewpoint, more important that to relate BP's properties to performance indicators, is identify and describe collective behaviors that help to maintain BP and perform passes, considering their relevance to the match.

For this topic, recent research has proposed several variables that compose collective movement behaviour (Memmert, Lemmink, & Sampaio, 2016b). When the analysis is focused on the dynamics of space occupation, variables such as the coverage area or effective playing space (Moura et al., 2012), length, width, and measures around the centroid (Folgado, Lemmink, Frencken, & Sampaio, 2014; Coutinho et al., 2019) are widely used. Besides that, several non-linear processing techniques have been used to improve the performance analysis process. For example, approximate entropy (ApEn) appears to provide information about the regularity of certain behaviour in soccer games and seems to be associated with adaptation during training interventions (Sampaio & Maçãs, 2012), critical moments of the game (Aguiar et al., 2017), or interpersonal game distances (Gonçalves et al., 2016). Complementarily to this structure of variability, the coefficient of variation (CV) has also been used to measure the magnitude of the

variability of a given behaviour across time (Gonçalves et al., 2017; Lorenzo-Martínez et al., 2019; Castillo, et al., 2019).

Non-linear processing techniques have also been used to identify coordination patterns in tactical behaviour analyses. Several studies have shown that movement synchronization is linked to tactical performance (Folgado, Duarte, Fernandes, & Sampaio, 2014; Folgado, Gonçalves, Sampaio, Folgado, & Gonçalves, 2017), with consequences on the external and internal workload demands (Folgado et al., 2018).

Considering the previous arguments, a multivariate approach based on metrics that describe collective behaviors in BP sequences could provide a more holistic model of this phenomenon in soccer matches. Within this topic, the outcomes would benefit from descriptions of the offensive and defensive game sequences to provide precise knowledge on the process. In addition, there are few studies on ball possession that describe collective tactical behaviours that determine the team ability to maintain ball possession. Thus, the purpose of this study was two-step: i) classify ball possession sequences according to the duration and number of passes; ii) identify which tactical variables most discriminate the different ball possession sequences, as classified in the previous step.

3.2. Methods

Data collection and sample

The Ethics Committee of the Campinas State University approved this research. The sample of this study corresponds to 527 ball possession (BP) sequences obtained from four first division official matches of the Brazilian Soccer Championship 2016.

The matches were recorded by two digital cameras (HDR-CX405, Sony), HD resolution, acquisition frequency of 15Hz, commonly used in collective tactical analysis (Rico-gonzález, Arcos, Nakamura, Arruda, & Pino-ortega, 2019). Subsequently, a semiautomatic tracking system was used to obtain the players' 2D positional data using the software DVideo (Pascual, Leite, & Barros, 2002; Figueroa, Leite, & Barros, 2006). The 2D coordinates of each player were defined as Xp(t) and Yp(t), where t represents each instant of time. The X and Y axes represent length and width of the pitch respectively. A Butterworth third-order low-pass digital filter with a cut-off frequency of 0.4 Hz was used as an external filter according to previous study recommendations

(Barros et al., 2007). DVideo software has an automatic tracking rate of 94% of the processed frames, an average error of 0.3 m for the determination of player position, and an average error of 1.4% for the distance covered (Figueroa, Leite, & Barros, 2006). Notational analysis was performed by an experienced operator to register the technical actions of each player, synchronized with the positioning data.

Ball possession sequences

Each ball possession started when any player controlled the ball through the successful execution of a technical action, such as a pass, interception or tackle, and restarting play, such as a free kick, throw-in, corner kick, and goal kick. When the game stopped for less than 15 seconds and the ball remained with the same team, it was considered the same BP sequence. This decision was made since the match dynamics of player positioning were not influenced. BP sequences of less than four seconds were excluded (to fulfil the nonlinear computation requirements). BP that did not contain at least one successful pass were also excluded.

Variables

Forty-one variables were computed and classified into three groups: notational, space occupation, and displacement synchronization (Table 3.1). Dynamic variables were analysed using the absolute values (mean), normalized approximate entropy (ApEn), and coefficient of variation (CV). ApEn is a nonlinear measure that quantifies the regularity in complex system behaviors (Pincus, 1991). For this study, we decided to compute the normalized entropy, a non-modified measure of regularity derived from the original ApEn, which is less dependent on time series length (Fonseca, Milho, Passos, Araújo, & Davids, 2012). Coefficient of variation (CV) values ((standard deviation/mean)×100) were used to verify the magnitude of variability of the time series.

The displacement synchronization variables consisted of the percentage of time that inter-player displacements were synchronized, calculated using the vector coding technique (Sparrow, Donovan, Van Emmerik, & Barry, 1987) and recently applied to investigate player behaviour during tennis matches (Pereira, van Emmerik, Misuta, Barros, & Moura, 2017). The technique consists of calculating the angle (ɵ) formed by the relative motion between two oscillators in two consecutive coordinates of a given time series. The coupling angle represents an instantaneous spatial relationship between two

players (dyad) in relation to the axes (X and Y). The coupling was considered as in-phase when the angle was at 45 ° or 225 ° (positive diagonal). Thus, the intervals 22.5 ° ≤ ɵ <67.5 ° and 202.5 ° ≤ ɵ <247.5° were chosen to assume an in-phase synchronization between two players. The synchronization percentage for each dyad was calculated for each team (in possession and without possession), in each ball possession sequence. The mean values of the percentage (% mean) of all the dyads were used to represent the mean of team synchronization and the CV (based on the % mean of all dyads) was calculated to indicate the variability between the dyads, i.e., if there was homogeneous behaviour of the team. All these procedures were performed for the X (longitudinal) and Y (lateral) axes of the pitch reference. Space occupation and synchronization variables are shown in Figures 3.1a and 3.1b, respectively. Data processing was performed in Matlab®2017(Mathworks Inc., Natick, MA, USA).

Table 3.1. Tactical variables used separated by groups.

Groups	Variables	Values
Notational	Time of possession	absolute value
	Successful pass	frequency
	Successful pass last third	frequency
	Shots	frequency
	Goal	frequency
Space occupation	Offensive team's effective playing space	mean, CV, ApEn
	Defensive team's effective playing space	mean, CV, ApEn
	Offensive team's length	mean, CV, ApEn
	Defensive team's length	mean, CV, ApEn
	Offensive team's width	mean, CV, ApEn
	Defensive team's width	mean, CV, ApEn
	Distance between offensive team's centroid and target	mean, CV, ApEn
	Distance between defensive team's centroid and target	mean, CV, ApEn
	Distance between team's centroid	mean, CV, ApEn
	Centroid Progression	absolute value
Displacement synchronization	Offensive team's synchronization X-axis	% mean, CV
	Defensive team's synchronization X-axis	% mean, CV
	Offensive team's synchronization Y-axis	% mean, CV
	Defensive team's synchronization Y-axis	% mean, CV

Forty-one variables were classified into three groups; notational (five variables), space occupation (twenty-eight variables), displacement synchronization (eight variables). Notational variables represent the total occurrence of the offensive team's ball possession, except time of possession. All continuous space occupation variables are calculated as mean, coefficient of variation (CV), and approximate entropy (ApEn) per ball possession per each team, except centroid progression that represents the difference between offensive team's centroid position in the last ball possession moment and the beginning of ball possession. For all displacement synchronization variables mean values of the percentage (% mean) of all the dyads were calculated to represent the mean of team synchronization and the CV was calculated to indicate the variability between the dyads. Abbreviations: CV = coefficient of variation; ApEn = approximate entropy; % mean = mean of the percentage.

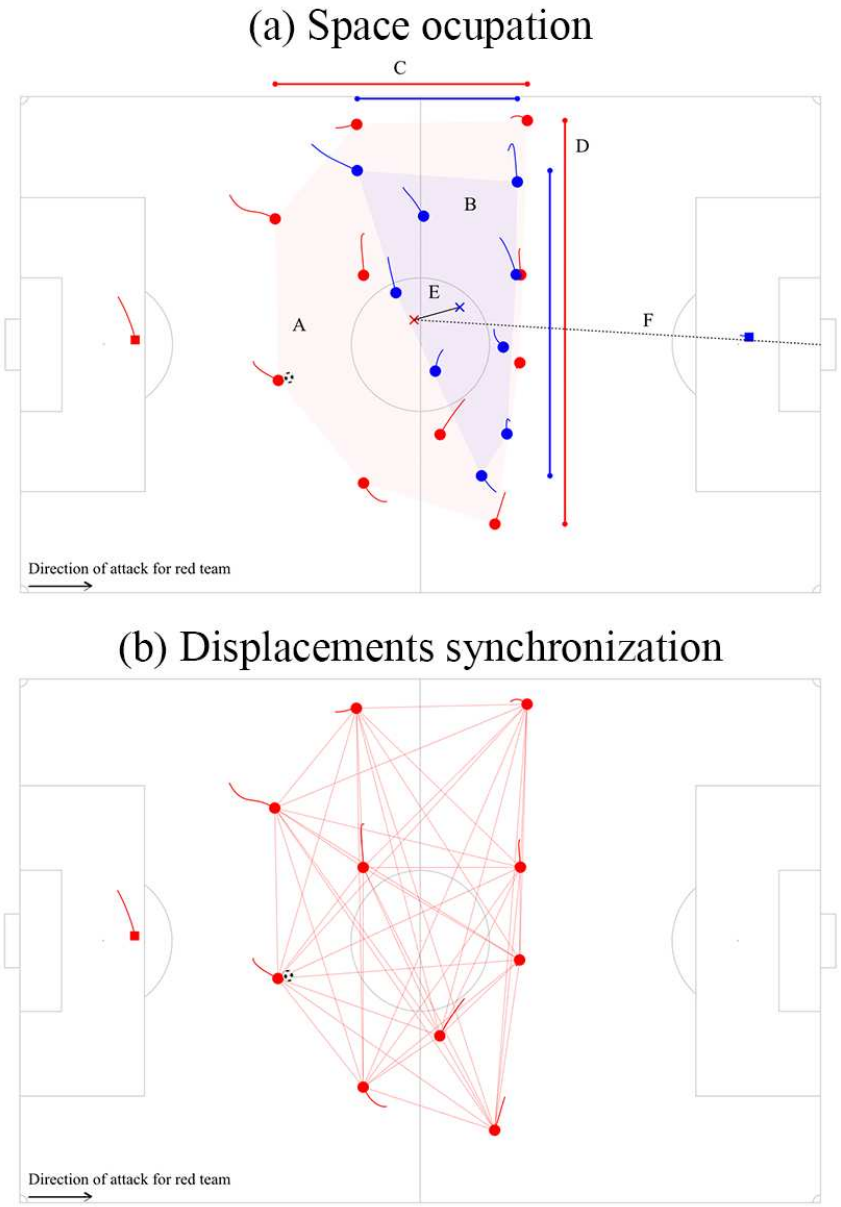


Figure 3.1. a) Representation of space occupation variables. Red team in ball possession (offensive phase) versus blue team (defensive phase) during long ball possession sequence. Abbreviations: A = Effective playing space (red team); B= Effective playing space (blue team); C = length (red team); D = width (red team); E = distance between team centroids; F = distance between centroid and target (red team). b) Representation of displacements synchronization. Each edge represents a dyad. Each player is connected to nine other players, except for the goalkeeper (total of 45 dyads).

Statistical analysis

A two-step cluster with log-likelihood as the distances measure and the Schwartz's Bayesian criterion was used to classify the ball possession sequences into the different groups, according to the time of possession and number of successful passes. Afterwards, a stepwise fisher´s discriminant analysis (FDA) was conducted to identify which variables best discriminate the previously obtained clusters. At each step, the variable that minimized the overall Wilks' Lambda was entered in the model. A minimum partial F (Fisher) value (3.84) to enter and maximum partial F value (2.71) to remove was used. Validation of discriminant models was conducted using the leave-one-out method of cross-validation. Was applied One-way ANOVA was used to compare the twelve selected variables into different groups (short, medium, and long ball possession sequences). Subsequently, the Bonferroni post-hoc test was utilized to identify pairwise differences. Statistical significance was set at 0.05 and the statistical analysis was carried out in IBM SPSS Statistics for Windows (Armonk, NY: IBM Corp). Complementarily, was observed the standardized mean differences and respective 95% confidence limits (CL), were also computed as magnitude of observed differences, effect size (Cohen´s *d*) and thresholds were: <0.2, trivial; 0.6, small; 1.20, moderate; 2.0, large; and >2.0, very large (Hopkins et al., 2009).

3.3. Results

The 527 ball possession sequences (BP) were classified into three different groups according to the time duration and number of successful passes: cluster 1 (short possessions n=295 or 55.8%, 11.07 ± 4.49s, 1.93 ± 0.99 successful passes), cluster 2 (medium possessions n=179 or 34%, 26.83 ± 7.33s, 5.41 ± 1.84 successful passes), and cluster 3 (long possessions n=53 or10.3%, 55.50 ± 14.97s, 12.11 ± 4.61 successful passes).

The stepwise fisher´s discriminant analysis (FDA) identified the most relevant variables to describe each cluster. The model consisted of two discriminant functions, with function 1 representing 95.8% of the total variance and function 2 representing 4.2%. The canonical correlations of functions 1 and 2 were, respectively, 0.83 and 0.30, with both functions being statistically significant (p <0.0001), (Wilks' Lambda = 0.27 and 0.91 for functions 1 and 2, respectively). The model presented a total of 81.6% of the original

grouped cases classified correctly. Table 3.2 presents the descriptive analysis for each variable, for the three clusters, as well as the structure coefficients (SC) for each function.

The variables that contributed most to the classification of the BP into function 1, in order of importance were: CV of the defensive synchronization-Y (SC = 0.58), CV of the defensive synchronization-X (SC = 0.42), successful pass last third, CV of the distance between offensive centroid and target (SC = 0.34), and mean of the offensive width (SC = 0.33). The remaining seven variables were: centroid progression, % mean of the offensive synchronization-X, CV of the offensive synchronization-X, % mean of the defensive synchronization-X, mean of the defensive length, and mean of the distance between offensive centroid and target.

Figure 3.2 represents the canonical discriminant function by distribution of the possession linked to cluster centroids, based on the discriminant scores represented by the X axis (function 1) and the Y axis (function 2).

Table 3.2. Descriptive and inferential statistics of different clusters of ball possession sequences.

Variables	Short (Mean ± SD)	Medium (Mean ± SD)	Long (Mean ± SD)	Short vs Medium (Mean difference ± CL) Effect size	Short vs Long (Mean difference ± CL) Effect size	Medium vs Long (Mean difference ± CL) Effect size	F1 95.8%	F2 4.2%
Time possession	11.07[ab] ± 4.49	26.83[c] ± 7.33	55.50 ± 14.97	15.76 ± 1.07 very large	44.43 ± 2.09 very large	28.67 ± 2.96 very large	-	-
Successful pass	1.93[ab] ± 0.99	5.41[c] ± 1.84	12.11 ± 4.61	3.49; ± 0.26 very large	10.19; ± 0.59 very large	6.70; ± 0.84 very large	-	-
CV DEF-SynY	47.25[ab] ± 13.55	30.11[c] ± 8.10	20.89 ± 4.51	-17.14; ± 2,19 large	-26.37; ± 3.70 very large	-9.23; ± 2.29 large	,577*	.244
CV DEF-SynX	40.60[ab] ± 14.43	27.00[c] ± 8.44	20.77 ± 5.64	-13.60; ± 2.33 moderate	-19.84; ± 3.95 large	-6.24; ± 2.43 moderate	.418*	.246
CV OFF-DCT	9.39[ab] ± 7.90	18.11[c] ± 9.83	20.55 ± 8.99	8.72; ± 1.62 moderate	11.16; ± 2.37 large	2.44; ± 2.97 small	-.346*	-.343
mean OFF-WID	40.41[ab] ± 6.84	45.97[c] ± 6.13	49.21 ± 4.62	5.56; ± 1.22 moderate	8.80; ± 1.92 large	3.24; ± 1.80 small	-.335*	-.112
CV OFF-SynX	44.99[ab] ± 16.03	35.10[c] ± 11.30	27.81 ± 7.64	-9.88; ± 2.69 moderate	-17.18; ± 4.42 moderate	-7.30; ± 3.26 moderate	.289*	.014
CProgress	12.72[b] ± 14.44[a]	22.02[c] ± 17.63	26.16 ± 13.07	9.30; ± 2.92 small	13.44; ± 4.18 moderate	4.13; ± 5.15 small	-.221*	-.136
mean OFF-DCT	55.00[ab] ± 14.63	47.56 ± 9.48	44.97 ± 8.51	-7.44; ± 2.41 small	-10.03; ± 4.08 moderate	-2.59; ± 2.86 small	.210*	.174
mean DEF-LEN	34.34[ab] ± 7.55	31.53[c] ± 7.00	27.58 ± 6.85	-2.80; ± 1.37 small	-6.75; ± 2.19 moderate	-3.95; ± 2.15 small	.190*	-.174
% mean OFF-SynX	47.37[ab] ± 15.85	42.52[c] ± 9.84	36.56 ± 6.20	-4.85; ± 2.58 small	-10.81; ± 4.35 moderate	-5.96; ± 2.82 moderate	.172*	-.119
% mean DEF-SynX	47.84[ab] ± 13.90	44.84 ± 10.47	42.85 ± 7.51	-3.00; ± 2.37 small	-4.99; ± 3.85 small	-4.03; ± 3.30 small	.097*	.018
Successful pass-LT	0.46[ab] ± 0.83	1.58[c] ± 1.68	3.38 ± 3.19	1.11; ± 0.22 moderate	2.92; ± 0.43 large	1.80; ± 0.65 moderate	-.381	.436*
% mean DEF-SynY	37.65 ± 11.62	37.16 ± 7.82	40.20 ± 6.57	-0.49; ± 1.92 trivial	2.54; ± 3.23 small	3.03; ± 2.33 small	-.025	.239*

Mean ± Standard deviation (SD), mean difference and respective 95% confidence limit (CL), effect size based on Cohen´s *d*, structure coefficient (SC) of 12 variables selected by the FDA model, and 2 variables used to separate the clusters (time of possession and successful pass). *variable better explained by function 1 or 2. One-way ANOVA and the Bonferroni post hoc to differentiate between groups (a = difference between clusters 1 and 2; b = difference between clusters 1 and 3; c = difference between clusters 2 and 3; $p<0.05$). Abbreviations: Short = Short ball possession sequences; Medium = Medium ball possession sequences; Long = Long ball possession sequences; F1 = Function 1; F2 = Function 2.

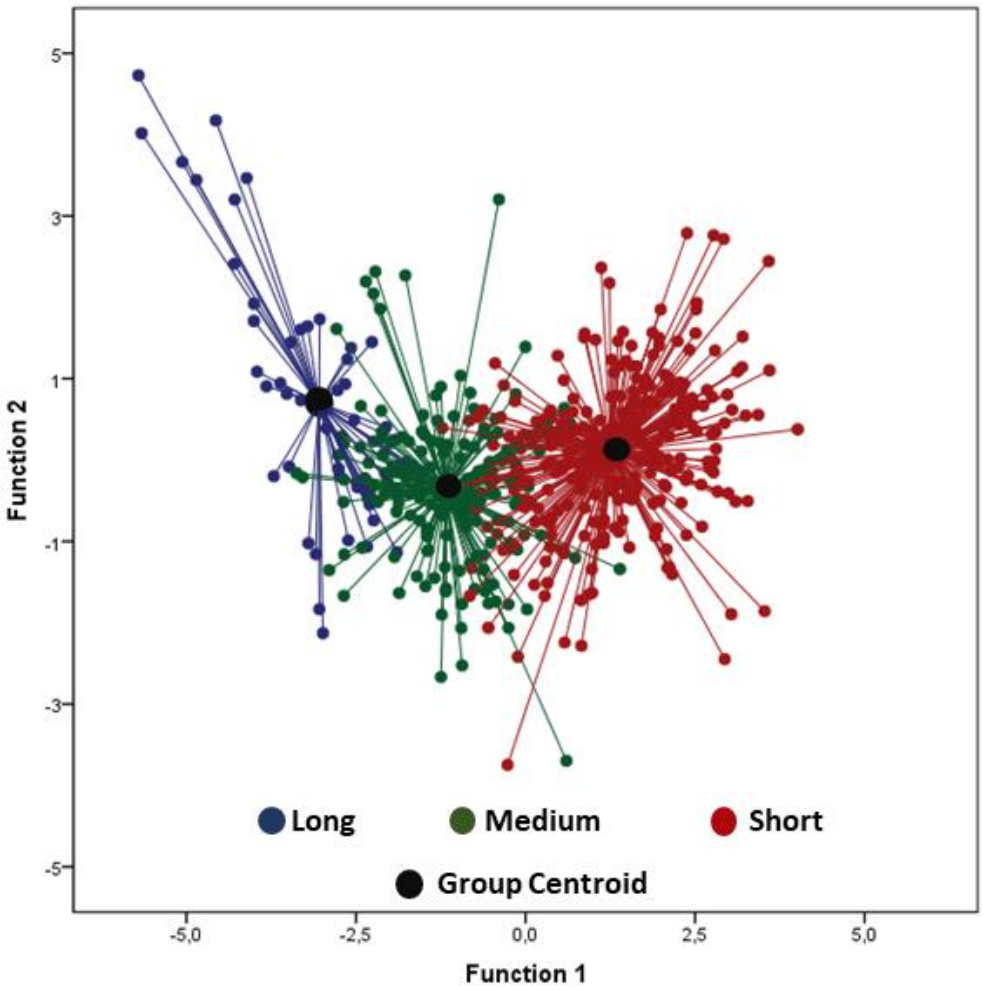


Figure 3.2. Territorial maps of the cluster centroids (group centroid) and their respective ball possession sequences (short = short ball possession; medium = medium ball possession; long = long ball possession) based on two canonical discriminant functions. Function 1 representing 95.8% of the total variance (0.83 of the canonical correlation) and function 2 representing 4.2% (0.30 of the canonical correlation), both functions being statistically significant ($p < 0.0001$), (Wilks' Lambda = 0.27 and 0.91 for functions 1 and 2, respectively).

3.4. Discussion

The purpose of this study was two-step: i) classify ball possession sequences according to the duration and number of passes; ii) identify which tactical variables most discriminate the different ball possession sequences, as classified in the previous step. In the first step, the cluster analysis classified the ball possession (BP) into three groups, short, medium and long duration. This classification allowed identify, describe and compare the collective tactical behavior to both teams, in offensive and defensive phase. For this, in the second step we use FDA to highlight, between forty-one tactical variables, the most relevant that better describe these three clusters. Five variables were highlighted: coefficient of variation (CV) of the defensive team's synchronization-Y, CV defensive team´s synchronization-X, successful pass last third, CV distance between offensive team's centroid and target, mean of the offensive team's width. The findings provided accurate tactical characterization to offensive and defensive team´s in the short, medium and long BP sequences and therefore suggest collective behaviors that help to maintain BP and perform passes, which is one of the challenges of the offensive phase of the matches.

In relation to the ball possession clusters identified, Aguiar et al. (2017) also classified BP using cluster analysis, however found two distinct groups, short and long, and the criterion for separation was based on centroid approximate entropy measurements. Jones et al. (2004) proposed three categories of ball possession durations, 3-10s, 10-20s, and more than 20s to investigate the relation with match status. Other studies with BP did not review the time duration or the number of passes and usually compared short and long sequences (Collet, 2013; da Mota et al., 2015; Yiannakos & Armatas, 2017).

In the present study, the short ball possession duration was characterized by lower successful passes in the last third, high CV of defensive team's synchronization in relation to X-axis and Y-axis, lower CV of distance between offensive team's centroid and target, and lower mean offensive team width. On the other hand, when we analysed the long ball possession duration, we observed more successful passes in the last third of the pitch, smaller CV of defensive team's synchronization in relation to X-axis and Y-axis, higher CV of distance between offensive team's centroid and target, and higher mean of the offensive team width. The medium ball possession duration presented intermediate values for the five variables.

The successful passes in the last third was the only notational variable highlighted. Displacement synchronization variables demonstrated importance for classification of the cluster, especially through the CV values of the defending team. These variables represent the variability of the percentage values of all team dyads. That is, the higher the CV, the more heterogenic the behaviour of the dyadic relations during the time series, as observed in short ball possessions. Otherwise, when dyads present similar behaviours between them, the CV values decrease, characterized in longer ball possessions. It is probable this behaviour is associated with the transition phases and stabilization in the possessions, i.e., when there is loss of the ball, the defensive team reorganizes strategically into its new tactical functions, changing the dynamics of space occupation during this transition. In short ball possessions, characterized as a mean of 11.7 s duration, there is no stabilization moment, or the transition phase is predominant, reflecting in the high CV of synchronization in relation to the X and Y axes. In the long possessions, there is also a transition phase, following a long period of stabilization, which probably explains the lower CV. These behaviours were conceptually identified by Hewitt et al. (2016), who generally describe the game as moments of frenetic attack to create imbalances in the opponent and moments of homeostasis, with rapid reorganization towards control and stability between the teams. Moura et al. (2013) also describe similar behaviour, but through the dynamics of the team occupying area, assigning higher values, based on spectral analysis, at the moment of the game where teams change ball possession rapidly, i.e., short possessions.

The other two highlighted variables belong to the 'space occupation' group. The CV of the distance between the offensive team's centroid and target indicated greater variability in longer ball possessions. It is probable the greater mobility of the team in possession exploring the pitch favoured the passes performed and control of the ball, as well as the width of the offensive team, which was higher in long ball possessions. It seems clear that teams adopting wider pitch space occupation and mobility favoured BP. Mobility and width are two of the five most important offensive principles proposed by Ouellette (2004). According to Clemente et al. (2013), the movements of players should extend to use the effective playing space by increasing the dispersion of players during the offensive phase. This behaviour makes it easier to attract defensive players to non-vital zones (e.g., lateral zones), thereby removing them from the vital zones (i.e., the middle zones). Clearly, it is essential to analyse offensive and defensive behaviour from

the interaction between teams, not just from a single perspective, as proposed by Fernandez-Navarro et al. (2016).

In summary, ball possession sequences were classified into three clusters based on the time possession and number of successful passes: short, medium and long duration. The discriminant analysis highlighted five most important variables to describe each cluster, and thus, these should be observed with more attention by coaches and sports scientists. Long ball possessions durations were characterized by more homogeneous behavior of the defending team in relation to displacements in lateral and longitudinal directions. There are few studies related to this phenomenon and therefore, their association with the micro-level relations among teammates should be further explored. Completely, higher width and mobility of the offensive team in long ball possession reinforcing some principles of offensive game advocated by experts, with the advantage of having been quantified and not only subjectively identified. This study used a limited sample based on Brazilian Soccer Championship and therefore should not be conclusive. The approach based on a multivariate model, using metrics recently proposed by research in performance analysis, allowed holistic analysis of the phenomena and provided accurate knowledge.

Session II

4. Session II (Study 2)

Classification and determinants of passing difficulty in soccer: a multivariate approach

ABSTRACT

Usually, the players' or teams' efficiency to perform passes is measured in terms of accuracy. The degree of difficulty of this action has been overlooked in the literature. The present study aimed to classify automatically the degree of passing difficulty in soccer matches and to identify and to discuss the variables that most explain the passing difficulty using spatiotemporal data. The data used corresponds to 465 passes, 32 independent variables and three classes of dependent variables. The Fisher Discriminant Analysis (FDA) presented 72.0% of the original grouped cases classified correctly. The passes analyzed were classified as low difficulty (56.5%), medium difficulty (22.6%), and high difficulty (20.9%). In general, high difficulty passes can be characterized as being associated with high pressure on the receiver; greater displacement and speed of the receiver; greater progression of the ball and rupture of opponents on the pitch; greater proximity to the opponent's goal; and fewer opponents between the receiver and the opponent's target. With less relevance, greater pressure on the passing player at the passing moment. Passes in soccer matches can be classified not only for their accuracy, but, based on their difficulty degree. The discriminant function coefficients presented allow to classify further datasets.

Keywords: passing; multivariate; soccer

4.1. Introduction

Tactics are the central component for the success in elite soccer (Rein & Memmert, 2016). Soccer matches have become more complex, faster, and players frequently need to work on reduced space to maintain ball possession (Wallace & Norton, 2014). In this context, the pass is the technical action most used to keep ball possession. On average, a typical match comprises 1,000 passes (Goes et al., 2018). This action, therefore, has been the main resource used to comply with the match offensive principles, i.e., maintain possession, progress in the pitch and create space and opportunity for scoring (Ouellette, 2004). In addition, it has been considered one of the key performance indicators (Cintia et al., 2015; F. Goes et al., 2019, 2018).

For these reasons, the pass has been investigated since Reep & Benajmin (1968), focusing on analyses based on frequency, density, and order of events (Chassy, 2013; Gyarmati, Kwak, & Rodriguez, 2014; M. Hughes & Franks, 2005; Lago & Martín, 2007; Mitschke & Milani, 2014; Peña & Navarro, 2015). Spatiotemporal data provided new perspectives to analyze pass actions. The accurate position of all players on the pitch allowed the proposal of new variables (Bush, Barnes, Archer, Hogg, & Bradley, 2015), metrics (Goes et al., 2018; Gyarmati & Stanojevic, 2016; Horton et al., 2014; Mchale, 2015; Power et al., 2017; Rein et al., 2017), indices (Cintia et al., 2015), and even predictions. Predictive modeling has explored different concepts, such as risk and advantage (Power et al., 2017), value of the passes (Spearman et al. 2017), quality of the pass (Horton & Gudmundsson, 2014), players' involvement in setting up goal-scoring chances by valuing the effectiveness of their passes (Bransen & Haaren, 2019).

The pass in soccer was defined as the deliberate act of touching and projecting the ball on de pitch to another teammate control over it, maintaining the possession of the team (Cunha et al., 2011; Horton et al., 2014; Wallace & Norton, 2014). When the ball reaches its intended destination, i.e., his/her teammates, the pass is considered successful.

We consider the pass as a technical-tactical action that occurs at time and space, in which the difficulty of the action depends on the interaction of several technical factors (e.g., body position and orientation, ball contact, movement speed, and pass distance) and tactical (e.g., team interaction and space occupation by individual players, group, or by the team), to the ball reaches its destination. Therefore, the passing difficulty refers to the degree of technical and tactical demands that the passing player has to complete the action successfully.

Usually, the players' or teams' efficiency to perform passes is measured in terms of accuracy, i.e., success rate of the passes, but the degree of difficulty of the action has been overlooked in the literature. The information of the degree of difficulty of each pass in the game would allow coaches to analyze the efficiency by relativizing the action difficulty. In addition, considering that the passing difficulty has a multivariate nature, it would be important to identify and discuss the variables that best explain this phenomenon. This would allow to extract more accurate information about an extremely frequent and important action for the success of the match.

The present study aimed to: (i) classify automatically the degree of passing difficulty in soccer matches; (ii) identify and discuss the variables that most explain the passing difficulty using spatiotemporal data. Our hypothesis is that the degree of passing difficulty depends on the technical and tactical variables combination associated with the passing player, receiver player, ball trajectory, and the pitch position where the action occurred.

4.2. Methods

Data collection and sample

The data used in this study corresponds to 465 passes randomly obtained from four matches of the first division Brazilian Football Championship 2016. Passes blocked and passes from corners and free kicks were not included in our analysis. The matches were recorded by two digital cameras Sony Handycam HDR-CX405, with HD resolution and acquisition frequency of 30 Hz. To obtain the players' 2D position data from the matches, we first sampled original data to 15Hz using the Virtual Dub software and then we used the software DVideo, which is a semiautomatic tracking system (Pascual, Leite, & Barros, 2002; Figueroa, Leite, & Barros, 2006). The players of each team were labeled as p = 1, 2, . . ., 14, including starting players and substitutes. Therefore, the 2D coordinates of each player (2D matrix) were defined as Xp(t) and Yp(t), where t represents each instant of time, while the X and Y axes represent length and width of the pitch respectively.

A Butterworth third-order low-pass digital filter with a cut-off frequency of 0.4 Hz was used as an external filter according to previous study recommendations (Barros et al., 2007). DVideo software has an automatic tracking rate of 94% of the processed frames, an average error of 0.3 m for the determination of player position, and an average error of 1.4% for the distance covered. After the filtering step, we used the DVideo

software to perform a notational analysis to register the technical actions synchronized with the players positional data. The Ethics Committee of the Campinas State University approved this research.

Variables

Thirty-two predictor variables (Table 4.1) were proposed for this study. For this purpose, three soccer experts, researchers, and coaches were interviewed separately and answered the following question: *"In your opinion, which information (technical and tactical actions) we can extract from the match is more relevant to determine the degree of passing difficulty in soccer?"* The answers were the basis for implementing the algorithm and for obtaining the predictor variables, in a two-dimensional (2D) perspective. Other variables proposed in similar previous studies (Horton et al., 2014; I. Mchale, 2015; Power et al., 2017; Rein et al., 2017) complemented the group of predictive variables to build a multi-class classification model. These variables were divided into groups and contributed as observation points for judgment (labeling process) by another group of experts. The observation points proposed were: a) pressure on the passing player; b) pressure on the passing receiver; c) ball trajectory; d) pitch position; and e) passing player techniques. All variables were obtained using the Matlab® *software*.

To evaluate pass degree, we have took into consideration two different moments: the origin of the pass (t_0), i.e., the exact moment of the contact with the ball by the passing player (PP); and destination of the pass (t_1), i.e., the exact moment of the contact with the ball in the subsequent action by the receiver player (RP), who may be his teammate (successful pass), or opposing team (unsuccessful pass). In both moments, we recorded the 2D positional information (XY) of the passing player ($PP_{(t0)}$) and the passing receiver player ($PR_{(t0)}$ and $PR_{(t1)}$), as well as all other players from both teams, team 1 (XY_1, XY_2,…, XY_{14}) and team 2 (XY_{15}, XY_{16}, ...,XY_{29}). We consider the pass as a vector ($\overrightarrow{AB}$) originating from $PP_{(t0)}$ (A) and ending in $PR_{(t1)}$ (B), projected on the pitch (Figure 4.1). Another vector, $\overrightarrow{AC}$, was based on the $PP_{(t0)}$ nearest opponent, i.e., with the origin in A and the extremity in the position nearest opponent (OP) to the passing player at t_0 moment, $OP_{(t0)}$ (C). The position variation of the PP also constituted an important vector, $\overrightarrow{DA}$, originating in $PP_{(t0-1)}$ (D) and extremity in $PP_{(t0)}$ (A).

In cases that the player did not perform a pass successfully (for instance, this pass was intercepted by an opponent) the position of the possible receiver of the pass (expected receiver - ER) was estimated according to the equation $ER = \frac{distance}{shortest\ distance} . \frac{angle}{shortest\ angle}$, as proposed by (Power et al., 2017). The ER position at the moment of the passing receipt, $ER_{(t1)}$, was used as $\overrightarrow{AB}$ vector extremity when passes were considered as an unsuccessful action and the calculation of other variables were based on the possible receiver position, both at t_0 and at t_1. This criterion was adopted considering that it is essential to observe characteristics of the PP intention to judge and determine its difficulty.

Table 4.1. Tactical variables used and abbreviations, separated by groups.

Groups	Abbreviation	Variables (description)
Passing player variables	Nearest opp. PP_{t0}	Distance between passing player and his nearest opponent at passing moment (t0).
	Density PP_{t0}	Number of opponents within the 1m, 2m, 5m and 10m radius to pass the player at t0.
	Velocity PP_{t0}	Instantaneous velocity of passing player at t0.
	Velocity nearest opp. PP_{t0}	Instantaneous velocity of nearest opponent to passing player at t0.
	Opponent angle	Angle (ɵ) between vectors $\overrightarrow{AB}$ and $\overrightarrow{AC}$ at t0. (cos ɵ = $\overrightarrow{AB}*\overrightarrow{AC}$ / $\lvert\overrightarrow{AB}\rvert*\lvert\overrightarrow{AC}\rvert$).
Passing receiver variables	Nearest opp. PR_{t0}	Nearest opponent to passing receiver player at t0.
	Density PR_{t0}	Number of opponents within the 1m, 2m, 5m and 10m radius to pass the receiver player at t0.
	Velocity PR_{t0}	Instantaneous velocity of passing receiver player at t0.
	Nearest opp. PR_{t1}	Nearest opponent to passing receiver player at t1.
	Density PR_{t1}	Number of opponents within the 1m, 2m, 5m and 10m radius to passing receiver at t1.
	Velocity PR_{t1}	Instantaneous velocity of passing receiver player at t1.
	Velocity nearest opp. PR_{t1}	Instantaneous velocity of nearest opponent to passing receiver player at t1.
	Displacement PR	Distance performed by passing receiver player between t0 and t1.
Ball trajectory variables	Passing distance	Passing distance (vector modules $\overrightarrow{AB}$).
	Passing angle	Angle (ɵ) between vector $\overrightarrow{AB}$ and unit vector $\vec{v}$ oriented by the X axis of the pitch (ɵ = arctan).
	Ball velocity	Mean velocity estimated by the ratio of the passing distance to the time between t0 and t1.
	Ball progression	Variation of the ball's position in relation to the X axis between t0 and t1.
	Outplayed opp.	Number of opponents between passing player at t0 and passing receiver player at t1 in relation X axis.
	Out ball angle	Angle (ɵ) between vectors $\overrightarrow{AB}$ and $\overrightarrow{DA}$. Calculation based on the angle between vectors (cos ɵ = $\overrightarrow{AB}*\overrightarrow{DA}$ / $\lvert\overrightarrow{AB}\rvert*\lvert\overrightarrow{DA}\rvert$).
Pitch position variables	Distance PP_{t0} to target	Distance between passing player and target of opponent at t0.
	Distance PR_{t0} to target	Distance between passing receiver and target of opponent at t0.
	Distance PR_{t1} to target	Distance between passing receiver and target of opponent at t1.
	Opp. btw PR_{t1} and target	Number of opponents between target and passing receiver player in relation X axis at t1.

Abbreviations: opp = opponent; PP_{t0} = passing player at the time of the pass execution; PR_{t0} = passing receiver at the time of the pass execution; PR_{t1} = passing receiver at the time of the receipt of the pass; btw = between.

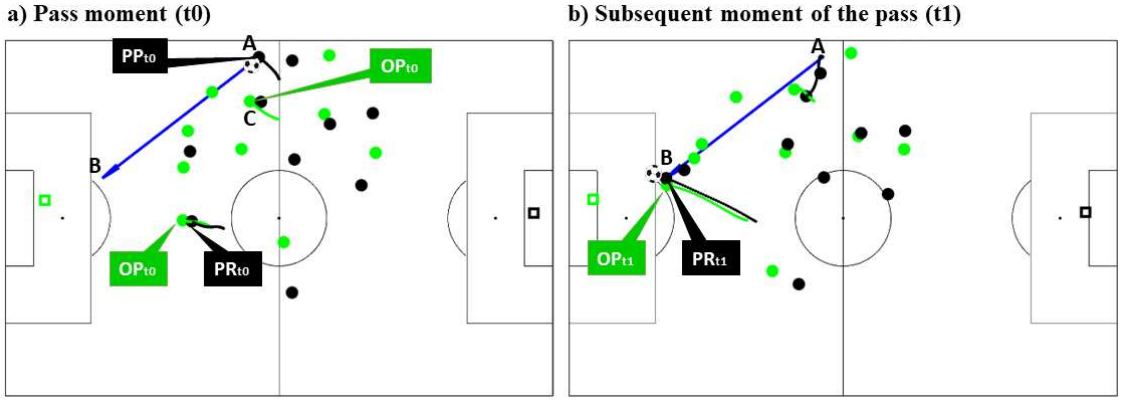


Figure 4.1. Illustration of the real pass situation, at the moment of contact with the ball (t0). PP_{t0} = passing player at the moment of the pass; PR_{t0} = receiver at the moment of the pass; OP_{t0} = nearest opponent to the passing player and receiver at the moment of the pass; A = origin of the pass; B = destination of the pass; C = OP_{t0} position. b) Illustration of the real pass situation at the moment of reception (t1). PR_{t1} = receiver at the moment of the reception of the pass. OP_{t1} = nearest opponent to the receiver when receiving the pass. Black team attacks to the left and gray team attacks to the right.

Labeling process

Two experts (researchers and coaches in soccer) performed, separately, the labeling process passes through judgment. Before judging the 465 passes, they were instructed about passing difficulty concepts, about points of observation, and were submitted to familiarization by watching examples of passes with different degrees of difficulty. For the purpose of this study, passing difficulty was defined as the degree of technical and tactical demands that the passing player must complete the action successfully. Then, they watched vide os of passes and assigned a classification for each event: class 1 (low difficulty), class 2 (medium difficulty), and class 3 (high difficulty). Experts could review the passes until they have a clear judgment. When they agreed about classification of the passes, the judgments were validated. When there was disagreement, a third expert decided about the classification. Only the classification of the first two experts was considered for the agreement test. The labels specified by the experts comprised the dependent variables. At the end of this process, we came up with a data set composed of 465 events (passes), 32 independent variables, and three classes of dependent variables (classes): $X = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x}n\}$, where $\underline{x}_i \in R^m$ and $m = 32$; and $Y = \{\underline{y}_1, \underline{y}_2, \ldots \underline{y}n\}$, where $y_i \in$ {low difficulty, medium difficulty, high difficulty}.

Statistical analysis

We adopted the use of the weighted kappa method (kw) to measure the inter-rater agreement between the experts (Cohen, 1968). A fisher's discriminant analysis (FDA) was used to classify the passes into three groups and identify which variables best discriminate them. Also, we used the leave-one-out cross-validation method to validate the proposed method. The interpretation of the obtained model took into consideration the Eigenvalue and structure coefficients (greater than |0.30|) that better distinguish the groups (Pedhazur & Manning, 1973).

Also, we use the One-way ANOVA method to compare sixteen variables selected into different classes (low, medium, and high difficulty pass). Finally, we used Tukey's post-hoc test considering a significance level of 5%. The statistical analyses were performed in the IBM SPSS Statistics for Windows (Armonk, NY: IBM Corp).

4.3. Results

We observed an inter-rater agreement between the experts of 80.2% in the labeling process, which corresponds to 373 events out of the 465 passes that comprise the data set used in this study. This result suggests a substantial agreement level (kw = 0.75) between the experts. The distributed into three classes considered in this study was 56.6% for the low difficulty passes (class 1), 22.6% for the medium difficulty passes (class 2), and 20.9% for the high difficulty passes (class 3). The FDA presented a total of 72.0% of the original grouped cases classified correctly. The percentage of successful passes within each class was 88.2% (low difficulty passes), 39.0% (medium difficulty passes), and 63.9% (high difficulty passes) (Figure 4.2).

Subsequently, the FDA was used to identify which variables most explain the passes classification in low, medium, and high difficulty. The model consisted of two discriminant functions, with function 1 representing 89.6% of the total variance and function 2 representing 10.4% (Figure 4.3). The canonical correlations of functions 1 and 2 were, respectively, 0.78 and 0.39, with both functions being statistically significant ($p < 0.0001$), (Wilks' Lambda = 0.32 and 0.84 for functions 1 and 2, respectively). The discriminant scores of the variables for each function are shown in Table 4.2.

The variables highlighted in function 1 in order of relevance based on structure coefficient (SC) were: Opponents between PR_{t1} and target, Density (5m) PRt0, Outplayed

opponents, Density (5m) PR_{t1}, Nearest opponent PR_{t1}, Nearest opponent PR_{t0}, Ball progress, Density (2m) PR_{t1}, Density (10m) PRt1, Velocity PR_{t1}, Density (10m) PR_{t0}, Displacement PR, Distance PR_{t1} to target. For function 2, the variables highlighted were: Nearest opponent PP, Density (10m) PP, Density (5m) PP. Table 4.2 presents the descriptive and inferential analysis for each variable, for the three classes, as well as the structure coefficients (SC) and discriminant function coefficients (FC) for each function.

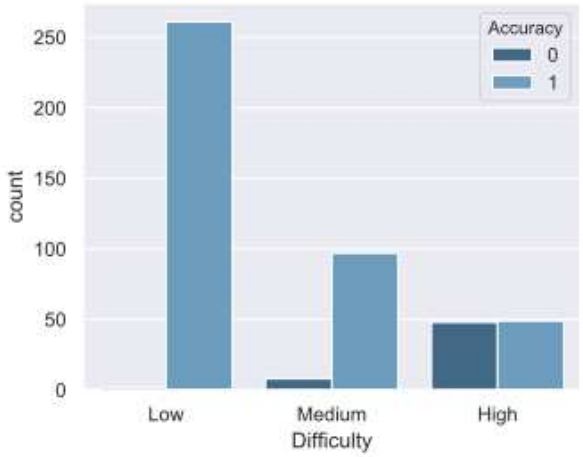


Figure 4.2. Distribution of 465 passes into three classes (low, medium, and high difficulty), and proportion of successful (1) and unsuccessful (0).

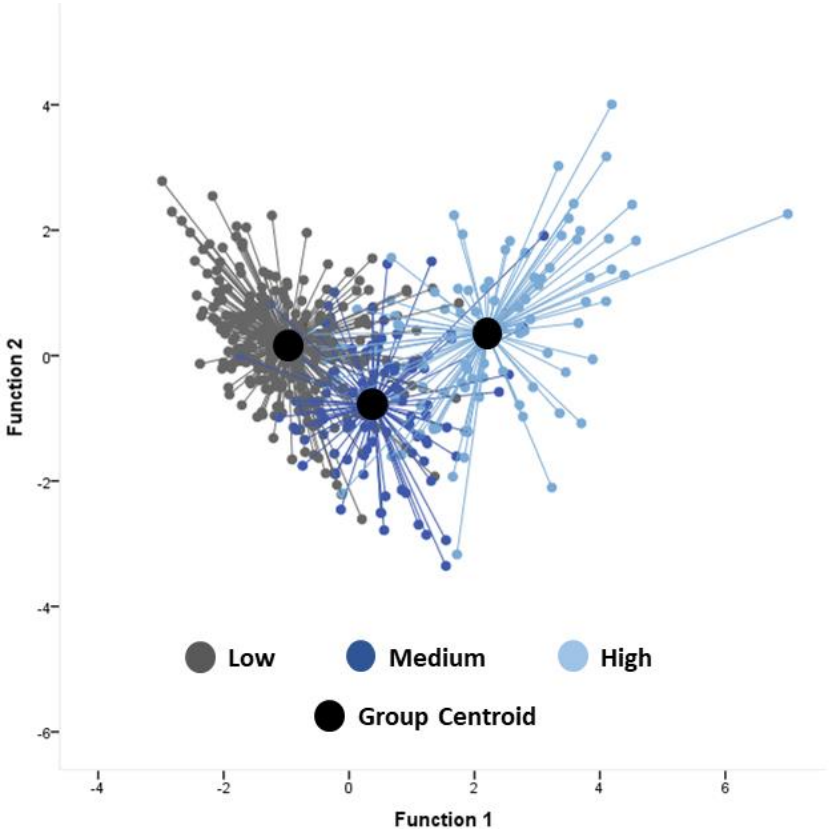


Figure 4.3. Territorial maps of the group centroid and their respective passes groups (low = low difficulty; medium = medium difficulty; long = long difficulty) based on two canonical discriminant functions.

Table 4.2. Descriptive and inferential statistics of three different classes (low, medium and high difficulty) of the passes.

	Variables	Low	Medium	High	F1 (SC)	F2 (SC)	F1 (FC)	F2 (FC)
		(Mean ± SD)	(Mean ± SD)	(Mean ± SD)	89.6%	10.4%	89.6%	10.4%
Pitch position variables	Opp. btw PRt1 and target	8.84[ab] ± 2.20	7.00[c] ± 2.32	4.90 ± 2.25	-0.562*	0.062	-0.227	0.340
	Distance PRt1 to target	56.14[b] ± 16.79	51.64[c] ± 15.62	37.84 ± 19.75	-0.324*	-0.190	-1.196	-2.696
Ball trajectory variables	Outplayed opponents	0.54[ab] ± 1.04	1.28[c] ± 1.69	2.82 ± 2.68	0.426*	0.143	0.180	0.534
	Ball progress	0.02[ab] ± 8.71	4.35[c] ± 11.43	12.82 ± 15.76	0.356*	0.102	-0.568	-0.641
Passing receiver variables	Density PRt0 (5m)	0.18[ab] ± 0.44	0.46[c] ± 0.57	1.08 ± 0.85	0.480*	0.188	0.316	0.245
	Density PRt1 (5m)	0.40[ab] ± 0.66	0.74[c] ± 0.69	1.35 ± 0.85	0.415*	0.089	0.105	0.239
	Nearest opponent PRt1	8.09[ab] ± 4.60	4.82[c] ± 3.23	3.16 ± 2.72	-0.406*	0.278	-0.026	0.277
	Nearest opponent PRt0	10.11[ab] ± 5.43	6.74[c] ± 4.39	4.06 ± 3.36	-0.403*	0.153	-0.226	-0.171
	Density PRt1 (2m)	0.04[ab] ± 0.20	0.19[c] ± 0.39	0.42 ± 0.52	0.354*	0.044	0.094	0.060
	Density PRt1 (10m)	1.36[ab] ± 1.23	2.11[c] ± 1.15	2.73 ± 1.42	0.353*	0.125	0.181	-0.338
	Velocity PRt1	7.34[ab] ± 4.97	11.04[c] ± 6.15	13.63 ± 7.30	0.352*	-0.165	0.251	-0.248
	Density PRt0 (10m)	1.09[ab] ± 1.24	1.59[c] ± 1.16	2.48 ± 1.58	0.334*	0.075	-0.087	-0.378
	Displacement PR	3.55[ab] ± 3.01	5.91[c] ± 5.69	8.48 ± 6.96	0.330*	-0.048	-0.188	-0.320
Passing player variables	Nearest opp. PP	6.02[ab] ± 4.23	3.19 ± 1.92	3.53 ± 2.56	-0.251	0.482*	-0.020	0.369
	Density PP (10m)	1.62[ab] ± 1.20	2.50 ± 1.17	2.30 ± 1.28	0.204	-0.463*	0.198	-0.301
	Density PP (5m)	0.67[ab] ± 0.75	1.18 ± 0.81	1.05 ± 0.74	0.186	-0.443*	0.087	-0.129

Mean ± standard deviation (SD), structure coefficient (SC), function coefficient (FC) of 16 variables selected by the FDA model. *Variable better explained by function 1 or 2. One-way ANOVA and the Bonferroni post hoc to differentiate between groups (a = difference between Low and Medium; b = difference between Low and High; c = difference between Medium and High; $p < 0.05$). Abbreviations: Opp = opponent.; F1 = Function 1; F2 = Function 2.

4.4. Discussion

The present study aimed to: (i) classify automatically the degree of passing difficulty in soccer matches; (ii) identify and discuss the variables that most explain the passing difficulty using spatiotemporal data. We identified that the most pass actions performed in the soccer matches were low difficulty passes, which correspond to 56.5% of pass actions considered in this study, followed by the medium difficulty passes that comprise 22.6%. Finally, the high difficulty passes comprise 20.9% of the passes. The FDA presented 72.0% of accuracy when classifying the degree of passing difficulty into three classes, lower than the experts' agreement, 80.21%.

Some recent similar studies also aimed to predict passes in soccer matches, however using concepts such as quality of the pass (Horton et al., 2014), risk of the pass (I. Mchale, 2015; Power et al., 2017; Spearman et al., 2017), value of the pass (Bransen & Haaren, 2019; F. Goes et al., 2018; Gyarmati & Stanojevic, 2016; Power et al., 2017; Rein et al., 2017). Risk represent the likelihood that the player will successfully make the pass and value the likelihood that the pass made will result in a shot within the next 10 seconds (Power et al., 2017). For classification model, Horton et al., (2014) obtained 85% accuracy to classifying passes as *good*, *ok*, *bad*, which essentially only give us the notion of quality of the passes, but the concept is not so clear.

The present study aimed to classify passes based on difficulty, that is, the degree of technical and tactical requirements that the player had to perform the pass: low, medium or high difficulty. This concept favors comparing successful and unsuccessful of the passes considering the difficulty of the actions. We found that 87.5% were classified as successful. When we analyze only high difficulty passes, the mean reduces to 50.5% (Figure 4.2). These numbers justifies the importance to analyze successful and unsuccessful passes relativizing by the difficulty of the action. Thus, the merit and ability of the player to perform passes with high difficulty are contemplated.

Another differential of this study was to highlight and discuss the variables that best explain the difficulty in performing passes and bring this information to a more applied context. Studies usually test variables to improve the accuracy of the prediction, but do not necessarily discuss the impact of each variable in the context of the game. In this study step, we identified 16 between 32 variables that best explain the degree of passing difficulty in soccer. These variables made it possible to quantitatively describe low, medium and high

difficulty passes and allow to classify further datasets with the discriminant function coefficients presented.

The FDA revealed through function 1 that the most important variables to determine the passing difficulty in soccer matches are related to the passing receiver, ball trajectory, and pitch position. In relation to the passing receiver, pressure variables at moment of the pass Density (5m and 10m) PR_{t0} and Nearest opponent PR_{t0} and at moment of the receipt, Density (2m, 5m, and 10m) PR_{t1} were highlighted. In addition, kinematic variables related to the displacement of the receiver, Displacement PR, and Velocity PR_{t1} were also highlighted. For the ball trajectory, function 1 highlighted variables that quantify the number of opponents won with the pass (outplayed opponents) and the progression of the ball in relation to the depth of the pitch (Ball progress). Besides that, two other highlighted variables, Opponents between PR_{t1} and target and Distance PR_{t1} to target represent, respectively, how many players there are between the receiver and the opposing target, and the position of the receiver when receiving the pass. Function 2, which explained only 10.4% of the variance, highlighted variables related to the pressure on the passing player at the time of the pass, Nearest opponent PP, and Density (5m and 10m) PP (Figure 4.4).

The results brought important considerations. First, the pressure variables on the passing receiver were more determinant than the pressure variables on the passing player. Pressure variables have been widely used in the literature, especially on-the-ball player in possession (Link et al., 2017, 2016), or in predictive passing studies (Horton et al., 2014). However, there was also a study that also highlighted distance to the passing receiver as an important variable to predict the quality of the pass (Horton et al., 2014).

Another important attention point was the variables related to the ball trajectory. It has been common to use angle and distance information from the pass to improve the level of information about this action (Bush et al., 2015; F. Goes et al., 2018). However, these variables did not have a relevant influence on the passing difficulty. The variable Ball progress, which synthesizes distance and orientation, highlighted the variables. In this group of variables, the variable Outplayed opponents, which represents the trajectory combined with the interaction between two teams, was the most important variable to determine the difficulty of pass actions. The variable Outplayed opponents was also an object of investigation in other studies (Rein et al., 2017; Steiner, 2018). It was observed that passes with origin in the middle third and destination in the offensive third won more opponents, and for this reason they are more effective and are related to the success in the matches (Rein et al., 2017). Finally, the most determining variable in function 1 was Opponents between

PRt1 and target. High difficulty passes have approximately five opponents between the receiver player and the target. The variable Outplayed opponents also represents the relationship of interaction between teams, which emphasize the importance of using a tracking system able to obtain data from both teams, such as multicamera systems.

In general, high difficulty passes can be characterized as high pressure on the receiver player at the passing moment (4.06 ± 3.36m), as well as at the receipt moment (3.16 ± 2.72m), greater displacement (8.48 ± 6.96m), and speed (13.63 ± 7.30 km / h) of the receiver between t0 and t1, greater progression of the ball (12.82 ± 15.76m) and rupture of opponents on the pitch (2.82 ± 2.68), greater proximity to the opponent's goal (37.84 ± 19.75 m), and fewer opponents between the receiver and the opponent's target (4.90 ± 2.25). With less relevance, greater pressure on the passing player at the passing moment (3.53 ± 2.56 m).

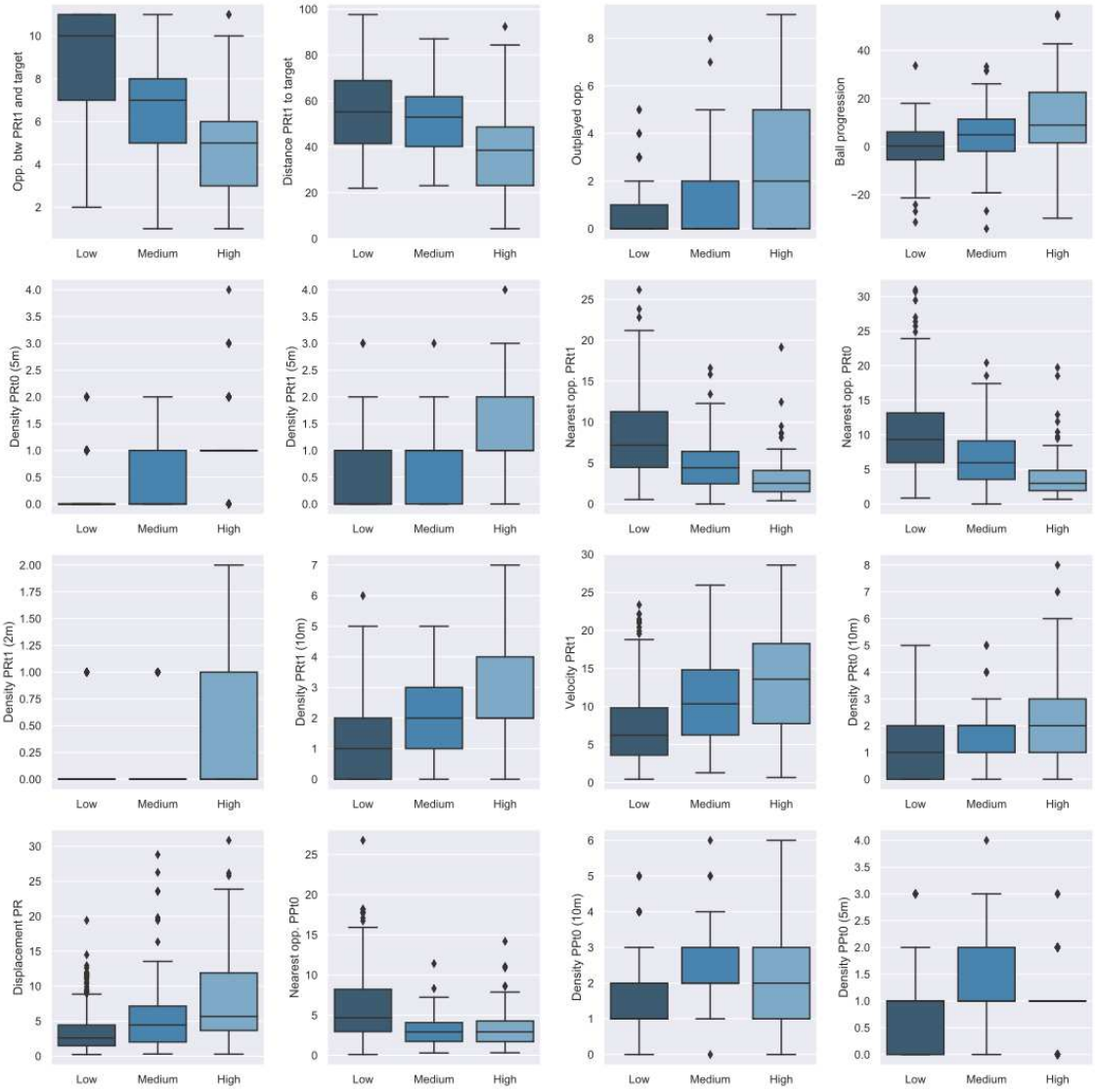


Figure 4.4. Comparison between three classes (low, medium, and high difficulty) of the passes for each sixteen variables highlighted by FDA.

4.5. Conclusions

The present study can contribute to a more accurate analysis of an extremely frequent and determinant action in soccer matches. Passes in soccer matches can be classified not only for their accuracy, but also based on their difficulty degree. The percentages of low, medium, and high passing difficulty suggest the match complexity or even the individual or collective quality when relating the passes accuracy with the different classes of difficulty. In addition, the highlighted variables should be looked at more carefully by coaches when analyzing the behavior of players, teams and offensive sequences. The values found for each variable can serve as a reference for planning training, such as small side games, and in future research.

Future research could focus on increasing the number of events, based on other competitive leagues, levels, age groups. In addition, the highlighted variables can help as a basis for other predictive models aiming at improving the accuracy in the classification of the passing difficulty in soccer matches.

Session III

5. Session III (Study 3)

Who are the best passing players in professional soccer? Machine learning approach classifies passes with different levels of difficulty and discriminate the best passing players

ABSTRACT

The efficiency of players and teams in performing passes is usually determined by their success rate. Classifying passes according to their difficulty level would allow the analysis of efficiency to be relativized by the ability of players and teams to execute high challenging passes. The present study aimed to classify automatically the level of passing difficulty in soccer matches using machine learning algorithms and to apply the model with the best performance to distinguish players and positions. We compared eight machine learning (ML) classifiers using 35 predictor technical-tactical variables based on spatiotemporal data. The Support Vector Machine (SVM) algorithm achieved the best performance with maximum balanced accuracy of 88.0%, considering the modeling of the target problem as a multi-class classification with three classes. In total, 2,522 pass actions were classified as low (53,9%), medium (23.6%), and high difficulty passes (22.5%). The percentage of successful passes for each class was 94.9%, 84.0%, and 49.3% for low (Low-DP), medium (Medium-DP), and high (High-DP) difficulty passes, respectively. The principal component analysis (PCA) showed a higher correlation between the accuracy in High-DP and Medium-DP with the first principal component (PC1). The PC1 scores were used to rank the best passing players. By analyzing the players' positions, we observed that the two best passing players were midfield and forward players. The proposed model improved the relevance of information of pass actions, which is the most frequent and determinant for performance in soccer matches.

Keywords: soccer; passing; match analysis; machine learning; team sports; player position.

5.1. Introduction

Analyzing soccer matches allows extracting information that favors the planning and direction of the training process in order to improve individual and collective performance. Historically, the analysis of technical demand in soccer matches, especially the passing, has focused on inferences using frequency, density, order, and accuracy of actions (Chassy, 2013; Gyarmati et al., 2014; M. Hughes & Franks, 2005; Lago & Martín, 2007; Mitschke & Milani, 2014; Peña & Navarro, 2015; Reep & Benajmin, 1968). This type of approach disregards the tactical aspects of the match. Tactics are the central component of success in elite soccer (Rein & Memmert, 2016).

Positional and time data of players have provided a more contextual analysis of the match. In addition, the increase of sports-related data available in terms of volume, velocity, and variety of data, the big data characterization (Riahi & Riahi, 2018), has required joint efforts from different areas, such as sports scientists and data scientists (Goes et al., 2020; Rein & Memmert, 2016). As a consequence, the application of machine learning (ML) and data mining (DM) techniques has increased considerably, with important contributions to performance analysis, injury prevention (Sikka et al., 2019), strategy analysis (Dick & Brefeld, 2019), training design, and talent identification (Goes et al., 2020).

Recently, the pass is one of the most investigated technical elements of a match, which is considered a key performance indicator in soccer analysis (Cintia, Giannotti, Pappalardo, Pedreschi, & Malvaldi, 2015; Goes et al., 2019). Some studies used machine learning techniques to predict passes based on concepts, such as risk and advantage of the passes (Power et al., 2017), time to intercept the ball from a pass (Spearman et al., 2017), quality of the passes (Horton & Gudmundsson, 2014), passing effectiveness and involvement of each player in creating score chances (Bransen & Haaren, 2019).

In our perspective, the pass is the basis of the soccer game. Soccer matches have become more complex, faster, and players frequently need to work on reduced space to maintain ball possession (Wallace & Norton, 2014). The pass is the most used action by the player in ball possession, representing 69% of the ball actions (Bransen & Haaren, 2019). On average, a typical match comprises 500 passes per team (Goes et al., 2018). Consequently, a player has the control of the ball to perform a pass every 10s, on average. In each pass, there will be a different context, with different levels of difficulty, influenced by technical and tactical factors, based on the strategy of both teams.

The efficiency of players and teams in performing passes is usually determined by their accuracy, i.e., the success rate of passes. This information does not reveal how complex the action was for the passing player. Classifying passes based on their level of difficulty would allow the analysis of efficiency to be relativized by the difficulty of the action, that is, the ability of players and teams to execute a high-level challenging pass.

We consider the pass as a technical-tactical action that occurs at time and space, in which the difficulty of the action depends on the interaction of several technical factors (e.g., body position and orientation, ball contact, movement speed, and pass distance) and tactical (e.g., team interaction and space occupation by individual players, group, or by the team), to the ball reaches its destination. Therefore, the pass difficulty refers to the degree of technical and tactical demands that the passing player must complete the action successfully. Accurate positional data over time of each player, of both teams, allows to represent these characteristics in a two-dimensional perspective. These variables may serve as the basis for a classification model to predict passes with different levels of difficulty. We believe that the classification of passes in different levels of difficulty could enable to distinguish players and position, taking into consideration the merit of successfully executing highly complex actions. This would bring important individual and collective performance indicators.

Therefore, the present study aimed to: (i) to classify automatically the level of passing difficulty in soccer matches using machine learning classifiers; (ii) to apply the model with the best performance to distinguish players and positions. Our hypothesis is that machine learning classifiers are effective to classify the level of passing difficulty based on technical and tactical variables combination, and that by classifying passes with different levels of difficulty we would be able to distinguish players and positions.

5.2. Methods

Study design

The present study consisted of five steps to build a classification model for automatically classifying pass actions according to their level of difficulty, which was used to predict a new sample. The steps are organized as follows (Figure 1): a) Data collection and sample; b) Predictor variables; c) Response variables (Labeling process); d) Dataset; e) Supervised

learning algorithms. After these steps, the model with the best performance was applied in different match analysis.

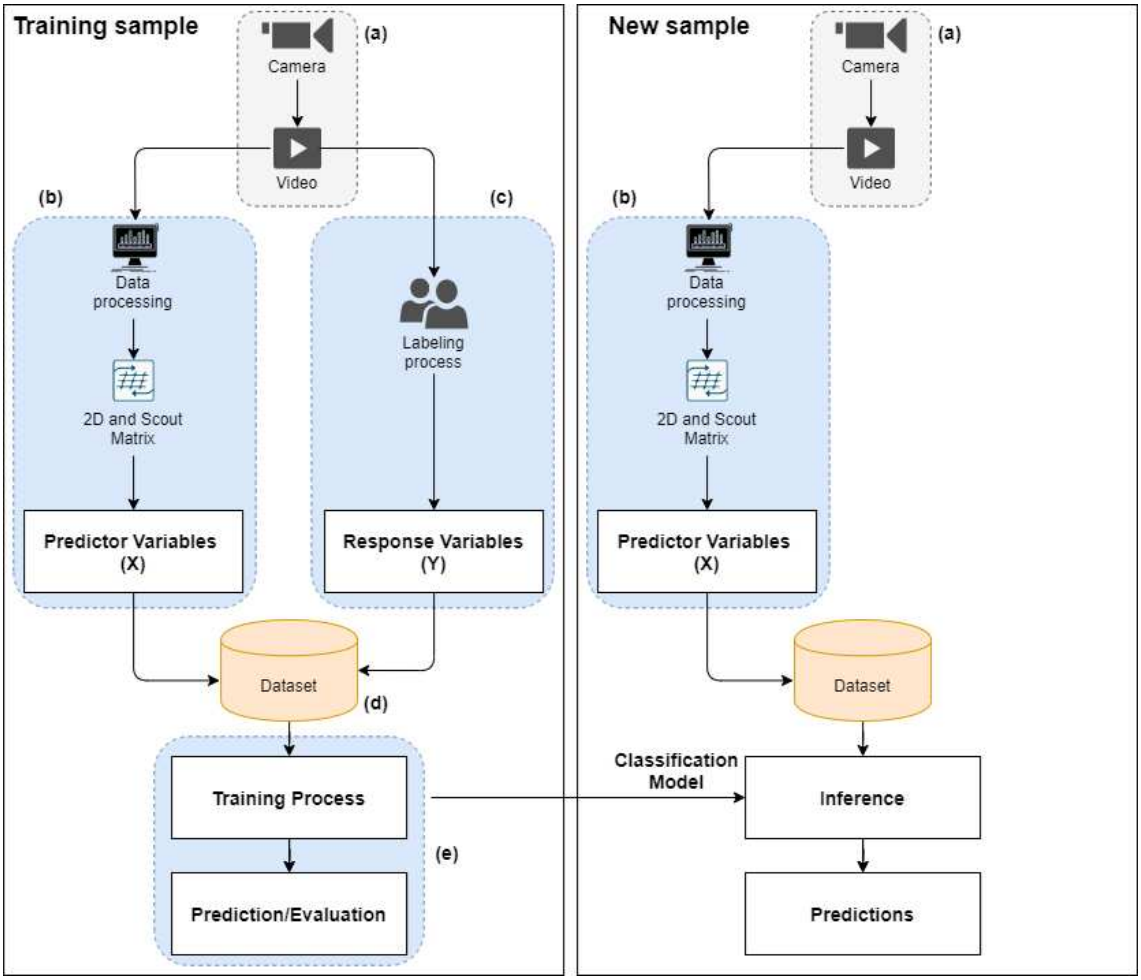


Figure 5.1: Study design starting from data collection (a), going through the processing to obtain the 2D and scout matrix, and consequently obtaining the predictor variables (b). Then, the labeling process was carried out to obtain the variables responses (c) until the composition of the dataset (d). From the dataset, the training process and evaluation of the algorithms were performed to obtain the best classification model (e). This process was performed with part of the sample (training sample). Finally, the model was applied to classify automatically passes in unseen samples.

Data collection and sample

The sample of this study comprised 2,522 passes obtained from four first division official matches of the Brazilian Football Championship 2016. The matches were recorded by two digital cameras Sony Handycam HDR-CX405, with acquisition frequency of 30 Hz. In order to reduce the amount of data to be processed, the videos were reduced to 15 Hz by Virtual Dub software. Subsequently, a semiautomatic tracking system was used to obtain the players' 2D positional data using the software DVideo (Pascual, Leite, & Barros, 2002; Figueroa,

Leite, & Barros, 2006). The players of each team were labeled as p = 1, 2, . . ., 14, including starting players and substitutes. Therefore, the 2D coordinates of each player (2D matrix) were defined as Xp(t) and Yp(t), where t represents each instant of time, and the X and Y axes represent length and width of the pitch respectively.

A Butterworth third-order low-pass digital filter with a cut-off frequency of 0.4 Hz was used as an external filter according to previous study recommendations (Barros et al., 2007). DVideo software has an automatic tracking rate of 94% of the processed frames, an average error of 0.3 m for the determination of player position, and an average error of 1.4% for the distance covered (Barros et al., 2007). After smoothing, notational analysis was performed by an experienced operator to register the technical actions, synchronized with the positioning data (Figueroa et al., 2006). The Ethics Committee of the University of Campinas approved this research.

Predictor variables

Thirty-six predictor variables ($\underline{x}_1, \underline{x}_2, \ldots \underline{x}n$) were proposed for this study (Table 1). For this purpose, three soccer experts, researchers, and coaches were interviewed separately and answered the following question: *"In your opinion, which information (technical and tactical actions) we can extract from the match is more relevant to determine the degree of passing difficulty in soccer?"* The answers were the basis to implement the algorithm and to obtain the predictor variables, in a two-dimensional (2D) perspective. Other variables proposed in similar previous studies (Horton, Gudmundsson, Chawla, & Estephan, 2014; Mchale, 2015; Power et al., 2017; Rein, Raabe, & Memmert, 2017) complemented the group of predictive variables to build a multi-class classification model. All variables were obtained using the Matlab®2017 *software*.

These variables were extracted from the 2D and scout matrix based mainly on two different moments: the origin of the pass (t_0), i.e., the exact moment of the contact with the ball by the passing player (PP); and destination of the pass (t_1), i.e., the exact moment of the contact with the ball in the subsequent action by the receiver player (RP), who may be a teammate (successful pass), or opposing team (unsuccessful pass). In both moments, we recorded the 2D positional information (XY) of the passing player ($PP_{(t0)}$) and the passing receiver player ($PR_{(t0)}$ and $PR_{(t1)}$), as well as all other players from both teams, team 1 (XY_1, XY_2,…, XY_{14}) and team 2 (XY_{15}, XY_{16}, ...,XY_{29}). We consider the pass as a vector ($\overrightarrow{AB}$) originating from $PP_{(t0)}$ (A) and ending in $PR_{(t1)}$ (B), projected on the pitch). Another vector,

$\overrightarrow{AC}$, was based on the $PP_{(t0)}$ nearest opponent, i.e., with the origin in A and the extremity in the position nearest opponent (OP) to the passing player at t_0 moment, $OP_{(t0)}$ (C). The position variation of the PP also constituted an important vector, $\overrightarrow{DA}$, originating in $PP_{(t0-1)}$ (D) and extremity in $PP_{(t0)}$ (A).

In cases that the player did not perform a pass successfully (for instance, this pass was intercepted by an opponent) the position of the possible receiver of the pass (expected receiver - ER) was estimated according to the equation $ER = \frac{distance}{shortest\ distance} . \frac{angle}{shortest\ angle}$, as proposed previously (Power et al., 2017). The ER position at the moment of the passing receipt, $ER_{(t1)}$, was used as $\overrightarrow{AB}$ vector extremity when passes were considered as an unsuccessful action and the calculation of other variables were based on the possible receiver position, both at t_0 and at t_1. This criterion was adopted considering that it is essential to observe characteristics of the PP intention to judge and determine its difficulty.

Table 5.1. Tactical variables used and abbreviations, separated by groups.

Groups	Abbreviation	Variables (description)
Passing player variables	Nearest opp. PP_{t0}	Distance between passing player and his nearest opponent at passing moment (t0).
	Density PP_{t0}	Number of opponents within the 1 m, 2 m, 5 m, and 10 m radius to pass the player at t0.
	Velocity PP_{t0}	Instantaneous velocity of passing player at t0.
	Velocity nearest opp. PP_{t0}	Instantaneous velocity of nearest opponent to passing player at t0.
	Opponent angle	Angle (ɵ) between vectors $\overrightarrow{AB}$ and $\overrightarrow{AC}$ at t0. (cos ɵ = $\overrightarrow{AB}$*$\overrightarrow{AC}$ / $\lvert\overrightarrow{AB}\rvert$*$\lvert\overrightarrow{AC}\rvert$).
	Foot /No foot	Indicates if the pass was performed with the foot or not (binary).
	One touth	Indicates if the pass was performed with the ball under the passer's previous control or not (binary).
Passing receiver variables	Nearest opp. PR_{t0}	Nearest opponent to passing receiver player at t0.
	Density PR_{t0}	Number of opponents within the 1 m, 2 m, 5 m and 10 m radius to pass the receiver player at t0.
	Velocity PR_{t0}	Instantaneous velocity of passing receiver player at t0.
	Nearest opp. PR_{t1}	Nearest opponent to passing receiver player at t1.
	Density PR_{t1}	Number of opponents within the 1 m, 2 m, 5 m, and 10 m radius to passing receiver at t1.
	Velocity PR_{t1}	Instantaneous velocity of passing receiver player at t1.
	Velocity nearest opp. PR_{t1}	Instantaneous velocity of nearest opponent to passing receiver player at t1.
	Displacement PR	Distance performed by passing receiver player between t0 and t1.
Ball trajectory variables	Passing distance	Passing distance (vector modules $\overrightarrow{AB}$).
	Passing angle	Angle (ɵ) between vector $\overrightarrow{AB}$ and unit vector $\vec{v}$ oriented by the X axis of the pitch (ɵ = arctan).
	Ball velocity	Mean velocity estimated by the ratio of the passing distance to the time between t0 and t1.
	Ball progression	Variation of the ball's position in relation to the X axis between t0 and t1.
	Outplayed opp.	Number of opponents between passing player at t0 and passing receiver player at t1 in relation X axis.
	Out ball angle	Angle (ɵ) between vectors $\overrightarrow{AB}$ and $\overrightarrow{DA}$. Calculation based on the angle between vectors (cos ɵ = $\overrightarrow{AB}$*$\overrightarrow{DA}$ / $\lvert\overrightarrow{AB}\rvert$*$\lvert\overrightarrow{DA}\rvert$).
	Passing angle	Angle(ɵ) btw vectors ($\overrightarrow{AB}$) e unit vector oriented by the X axis of the pitch. (ɵ = arctan) (categorical).
	Passing accuracy	Indicates pass success and failure (binary).
Pitch position variables	Distance PP_{t0} to target	Distance btw passing player and target of opponent at t0.
	Distance PR_{t0} to target	Distance btw passing receiver and target of opponent at t0.
	Distance PR_{t1} to target	Distance btw passing receiver and target of opponent at t1.
	Opp. btw PR_{t1} and target	Number of opponents between target and passing receiver player in relation X axis at t1.

Abbreviations: opp = opponent; PP_{t0} = passing player at the time of the pass execution; PR_{t0} = passing receiver at the time of the pass execution; PR_{t1} = passing receiver at the time of the receipt of the pass; btw = between.

Response variables (Labeling process)

Approximately 20% of the total samples (n = 465 passes) were randomly separated for the passes labeling process. Two experts (researchers and coaches in soccer) performed, separately, the labeling process passes through judgment. Before judging the 465 passes, they were instructed about passing difficulty concepts and were submitted to familiarization by watching examples of passes with different degrees of difficulty. For this study, passing difficulty was defined as the degree of technical and tactical demands that the passing player must complete the action successfully. Then, they watched videos of passes and assigned a classification for each event: class 1 (very low difficulty), class 2 (low difficulty), class 3 (medium difficulty), class 4 (high difficulty), and class 5 (very high difficulty). Experts could review the passes until they have clear judgment. When they agreed about the classification of the passes, the judgments were validated. When there was disagreement, a third expert decided about the classification. Only the classification of the first two experts was considered for the agreement test. The labels specified by the experts comprised the dependent variables of the model: $Y = \left\{\underline{y}_1, \underline{y}_2, \ldots \underline{y}n\right\}$.

Dataset

The obtained predictor variables (X) and response variables (Y) were used to fit a supervised classification model. We modeled the problem considering five classes, as described in the previous section, and three classes, by joining two extreme classes (very low difficulty and low difficulty; high difficulty, very high difficulty), and maintaining the intermediate class (medium difficulty). Thus, the dataset structure was composed of 465 events (passes), 35 predictor variables, and two options of response variables, with five classes (condition 1) and three classes (condition 2):

$X = \left\{\underline{x}_1, \underline{x}_2, \ldots \underline{x}n\right\}$ where $\underline{x}_i \in R^m$ and $m = 35$;

$Y = \left\{\underline{y}_1, \underline{y}_2, \ldots \underline{y}n\right\}$ where $y_i \in$ {very low difficulty, low difficulty, medium difficulty, high difficulty, very high difficulty} or $y_i \in$ {low difficulty, medium difficulty, high difficulty}.

Supervised Learning Classifiers

We compared eighth supervised learning classifiers available in the scikit-learn v0.20.3 library (Pedregosa, Weiss, & Brucher, 2011), and traditionally used in classification problems: Random Forest (RF), Logistic Regression (LR), K-Neighbors Nearest (K-NN), Support Vector Machine (SVM), Linear Support Vector Machine (LSVM), Gaussian Naïve

Byes (NB), Linear Discriminant Analysis (LDA) and MLP (Multilayer Perceptron). A brief description and key characteristics of the classifiers are presented in Table 2.

Table 5.2. Background of Machine Learning Classifiers

Classifiers	**Characteristics**
RF	Ensemble of classifiers that use many decision tree models, the predictions of which are combined by majority voting to produce a single output (Montoliu, Martín-félez, & Torres-sospedra, 2015).
LR	Based on the probability for a sample to belong to a class. The probabilities must be continuous in R and bounded between (0, 1). The name logistic comes from the sigmoid (or logistic) function (Bonaccorso, 2017).
K-NN	Consists of assigning a new test sample to the class most frequently represented among the k closest instances in the training set according to a certain dissimilarity measure (Montoliu et al., 2015).
SVM	Use a kernel to transform the original data into a higher dimensional space, where the hyperplane that optimally separates the data into two categories is found. We are using SVM to refer to the mapping functions based on a Radial Basis Function (RBF) (Montoliu et al., 2015).
LSVM	We are using LSVM (Linear SVM) to refer to the mapping functions based on a Linear Function (LF) (Montoliu et al., 2015).
NB	A family of powerful and easy-to-train classifiers that determine the probability of an outcome given a set of conditions using Bayes' theorem, i.e., the conditional probabilities are inverted, so that the query can be expressed as a function of measurable quantities (Bonaccorso, 2017).
LDA	The decision boundaries created by LDA are linear, leading to decision rules that are simple to describe and implement. A new observation is classified to the class with closest centroid, based on the Mahalanobis metric, using a pooled covariance estimate (Hastie, Tibshirani, & Friedman, 2008).
MLP	Non-linear feedforward artificial neural network which consists of weighted interconnected layers of computational units (neurons) in a directed graph (Montoliu et al., 2015).

Abbreviations: Random Forest (RF), Logistic Regression (LR), K-Neighbors Nearest (K-NN), Support Vector Machine (SVM), Linear Support Vector Machine (LSVM), Gaussian Naïve Byes (NB), Linear Discriminant Analysis (LDA) and MLP (Multilayer Perceptron).

The data were processed in a *python 3.6* environment, following some steps until obtaining the classifiers performance indicators:

Pre-processing: We organized the dataset into predictor variables (X) and response variables (Y) and we split the dataset into two subsets: training set (75%), and test set (25%). We also scale the predictor variables by applying the Z-score normalization.

Evaluation protocol: The training set was used to determine the curve with the best fit and the grid search was applied to obtain the best parameters considering a k-fold cross-validation protocol (k = 5). We used the test set only to evaluate the classifiers' performance.

Evaluation metrics: We adopted the use of balanced accuracy and f1-score to measure the performance of classifiers. We repeated the experiment ten times considering different seeds in order to measure aspects of generalization of the models. With this, we end up with 50 values of balanced accuracy and f1-score, i.e., ten values for each one of the five rounds of cross-validation protocol.

Application of model to Match Analysis

The classification model with the best performance was used to predict a set of unseen samples of passes (n = 2.057). The predicted sample plus the previously labeled sample comprised 2,522 passes. These 2,522 passes were classified according to the degree of difficulty and were therefore used to make inferences in the four games used in this study. Two main inferences were made: analysis of players and positions.

Firstly, the players were categorized into six roles: GK (goalkeeper), external defenders (ED), central defenders (CD), defensive midfield (DM), offensive midfield (OM), and forwards (FW). Thus, the goalkeeper was defined as the player with the lowest average for the x-axis (goal-line), while the right and left external defenders with the minimum and maximum average for the y-axis, respectively, but with x-axis values less than 55 m. The other roles were defined using the k-medoids algorithm, considering four clusters and the squared Euclidean distance. We developed a python-based tool to extract information from datasheets to obtain the average position of each player. This positioning data fed our algorithm, implemented in Matlab®, that defines the players' roles. From that, seventy-seven players were analyzed during four games end categorized as: 5 GK, 12 ED, 13 CD, 9 DM, 23 OM, 15 FW. Subsequently we used principal component analysis (PCA) to identify players with better performance in performing passes with different degrees of difficulty. For

this purpose, players who did not perform at least 20 passes in total or at least five passes in each class were excluded.

Statistical analysis

Firstly, we adopted the use of the weighted kappa method (kw) to measure the inter-rater agreement between the experts (Cohen, 1968). In the first part of this study, to compare the classifiers performance, we used Friedman test based on the average and standard deviation values of the balanced accuracy, k-fold (n = 5). The test was replicated ten times totaling 50 balanced accuracy values. When there was rejection of the null hypothesis, that is, equality between classifiers, a Nemenyi post hoc test was used to identify the differences. P value was used for comparison between all pairs. A p-value below 0.05 indicates that that comparison is statistically significant, that is, it is unlikely that two sets for error rates are samples from the same distribution. This step was performed in a *python 3.6* environment (library), and based on the proposal by (Demsar, 2006) which suggests the use of non-parametric tests, especially those used in this study, for multiple comparison of machine learning data.

In the second part of this study, we use principal component analysis (PCA) to identify the best passing player based on the percentage of successful passes in low, medium and high classes. We used the PCA for three reasons: to explain the variability between the variable accuracy in low, medium, and high difficulty passes among 41 players analyzed. To identify which of these three variables are most determinant to distinguish the best passing players and positions; and to rank the best passing players and positions. The explained variance was based on the eigenvalues of each component. The correlation of the variables and each of the principal components was observed by the component matrix. The ranking of the best passing players was based on the scores of the first principal component (PC1). The PCA analysis was processed using IBM SPSS Statistics for Windows (Armonk, NY: IBM Corp).

5.3. Results

The results obtained in the present study are presented according to the following sequence: characterization of the labeled sample (n = 465), comparison of the classifiers' performance and application of the model with the best performance for match analysis.

The 465 passes were initially labeled into five classes by the experts, condition 1, and later reduced to three classes, condition 2. We observed an inter-rater agreement between the experts of 65.6% in the labeling process, which corresponds to 305 events out of the 465 passes, condition 1, and inter-rater agreement between the experts of 80.2%, which corresponds to 373 events out of the 465 passes, condition 2. This result suggests a substantial agreement level, *kw* = 0.73 and 0.75, between the two experts for both condition, 1 and 2 respectively. The distributed into five classes was 26.0% for the low difficulty passes (class 1), 30.5% for the low difficulty passes (class 2), 22.6% for the medium difficulty passes (class 3), 15.1% for the high difficulty passes (class 4), and 5.8% for the very high difficulty passes (class 5), Figure 2a. For the three classes, the distribution was 56.6% for the low difficulty passes (class 1), 22.6% for the medium difficulty passes (class 2), and 20.9% for the high difficulty passes (class 3), Figure 2b.

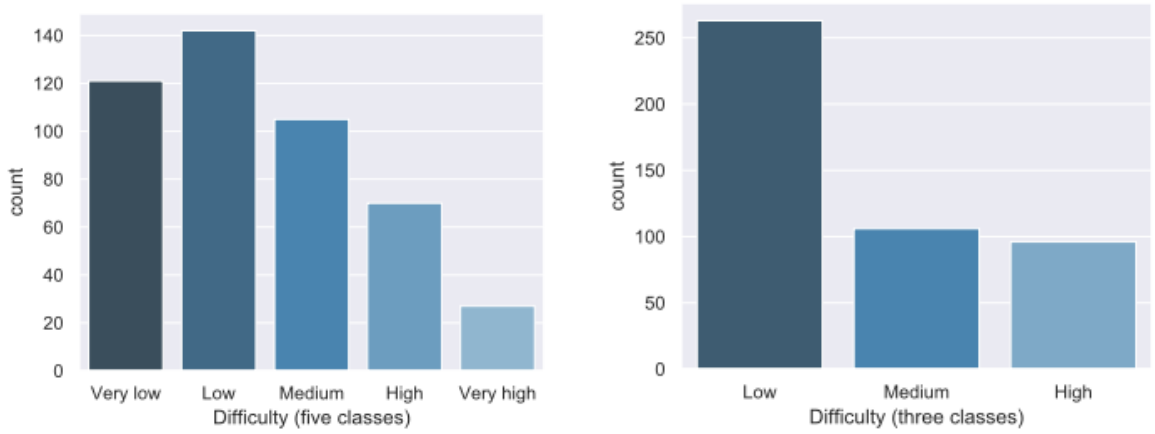


Figure 5.2. a) Distribution of 465 passes in five classes according to experts in the labeling process. b) Distribution of 465 passes in three classes according to experts labeling process.

Subsequently, the labeled dataset was used to train supervised learning classifiers. The two conditions, five and three classes were tested for the eight classifiers, totaling 16 classification models. Table 3 summarizes the performance of the classifiers in both conditions considering values of balanced accuracy and F1-score. For five classes, the best performance classifiers were LDA (0.58 ± 0.10), SVM (0.56 ± 0.10) and LR (0.56 ± 0.08), Figure 3. For three classes, the balanced accuracy values were higher in relation to the five classes for all classifiers. The best performances for three classes were SVM (0.70 ± 0.04), LR (0.70 ± 0.05), and LDA (0.68 ± 0.05), that presented statistical difference for the others (Figure 4). In addition, SVM (0.71 ± 0.08) e LR (0.73 ± 0.07) presented higher F1-score values for the other classifiers, and there was no statistical difference between them. Among

all the classifiers analyzed, we chose to choose the SVM which, although there was no significant difference for the LR and LDA, was the one that reached the highest balanced accuracy value (0.88) in one of the rounds, that is, 88% correct when automatically classifying passes into three classes, low, medium, and high difficulty. The confusion matrix of the chosen model can be seen in Figure 5. We adopted the balanced accuracy values, that is, average percentage of correctness by classes based on k-fold cross validation. Our choice was based on the nature of the problem, that focused on the model's ability to correctly classify as many events as possible.

Table 5.3. Comparison of performance between Machine Learning Classifiers.

Classes	Metrics	SVM	LR	LDA	L-SVM	MLP	NB	RF	K-NN
Three	Bal. Acc.	0.70 ± 0.04	0.70 ± 0.04	0.68 ± 0.05	0.67 ± 0.04	0.64 ± 0.03	0.62 ± 0.03	0.62 ± 0.03	0.58 ± 0.05
	Best Acc.	0.88	0.87	0.80	0.78	0.85	0.75	0.74	0.82
	F1-score	0.71 ± 0.08	0.73 ± 0.07	0.75 ± 0.07	0.73 ± 0.07	0.72 ± 0.06	0.71 ± 0.07	0.72 ± 0.05	0.70 ± 0.07
Five	Bal. Acc.	0.56 ± 0.10	0.56 ± 0.08	0.58 ± 0.10	0.53 ± 0.09	0.55 ± 0.09	0.42 ± 0.08	0.51 ± 0.10	0.48 ± 0.09
	Best Acc.	0.75	0.70	0.75	0.71	0.76	0.58	0.74	0.65
	F1-score	0.59 ± 0.08	0.58 ± 0.06	0.59 ± 0.08	0.54 ± 0.06	0.56 ± 0.07	0.41 ± 0.06	0.56 ± 0.08	0.53 ± 0.07

Abbreviations: RF = Random Forest, LR = Logistic Regression, K-NN = K-Neighbors Nearest, SVM = Support Vector Machine, LSVM = Linear Support Vector Machine, NB = Gaussian Naïve Byes, LDA = Linear Discriminant Analysis and MLP = Multilayer Perceptron, Bal. Acc. = Balanced Accuracy, Best Acc. = Best Accuracy.

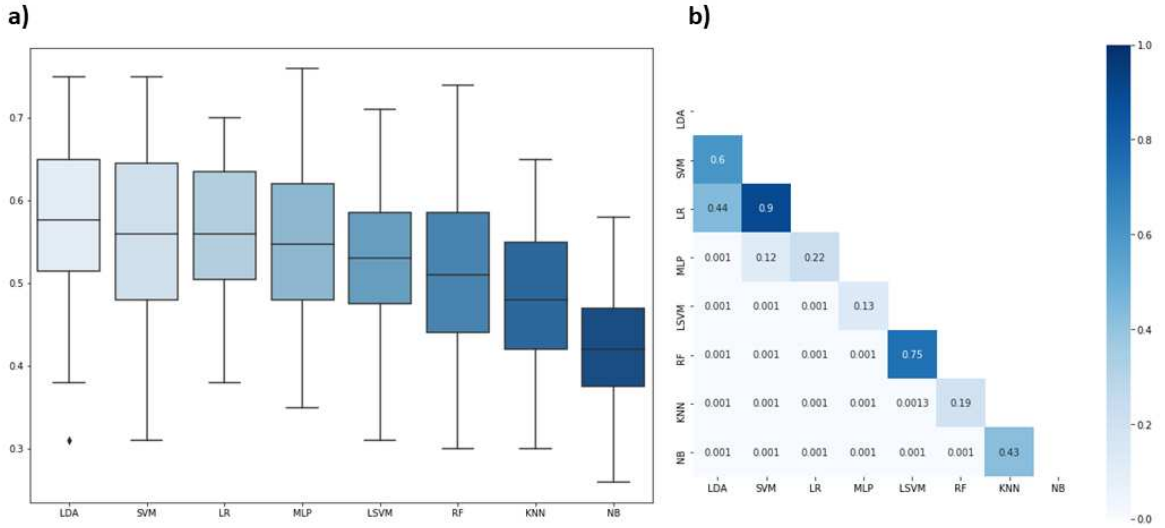


Figure 5.3. a) Comparison of performance based on balanced accuracy between machine learning classification in the condition 1 (five classes) using boxplot and Friedman statistical test. b) Pairwise comparison. The scale represents the p-value obtained through the Nemenyi post hoc test, also indicated into the squares (p-value below 0.05 indicates a statistically significant difference). Legend: LDA = Linear Discriminant Analysis, SVM = Support Vector Machine, LR = Logistic Regression, MLP = Multilayer Perceptron, LSVM = Linear Support Vector Machine, RF = Random Forest, K-NN = K-Neighbors Nearest, NB = Gaussian Naïve Byes.

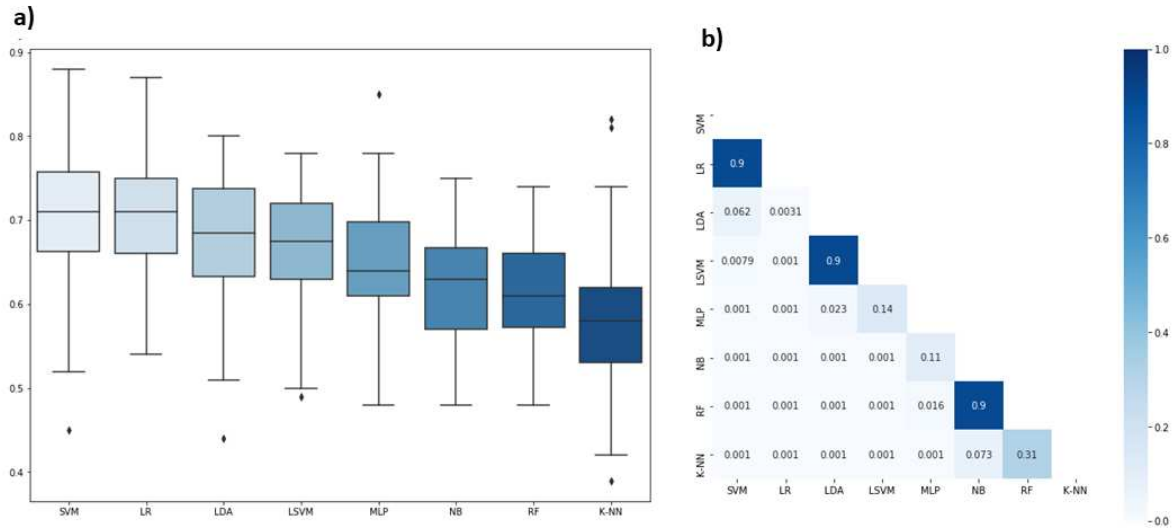


Figure 5.4. a) Comparison of performance based on balanced accuracy between machine learning classification in the condition 1 (three classes) using boxplot and Friedman statistical test. b) Pairwise comparison. The scale represents the p-value obtained through the Nemenyi post hoc test, also indicated into the squares (p-value below 0.05 indicates that that comparison is statistically significant). Abbreviation: SVM = Support Vector Machine, LR = Logistic Regression, LDA = Linear Discriminant Analysis, LSVM =

Linear Support Vector Machine, MLP = Multilayer Perceptron, NB = Gaussian Naïve Byes, RF = Random Forest, K-NN = K-Neighbors Nearest.

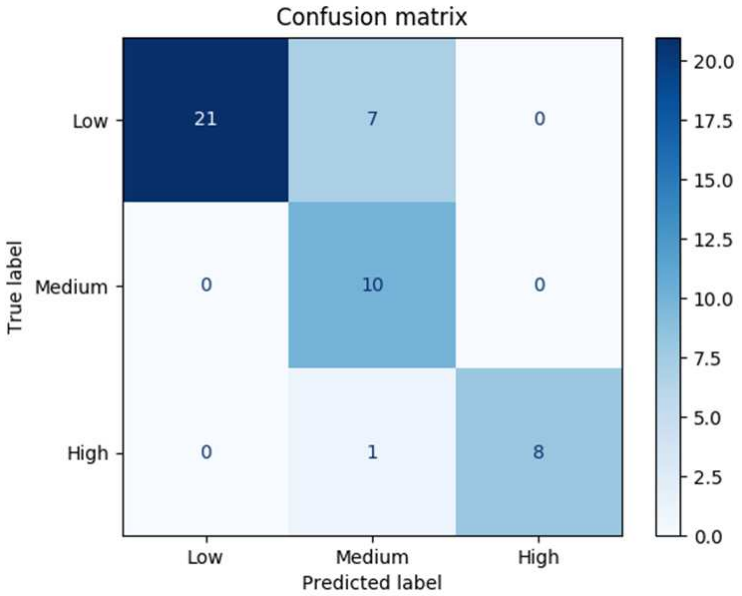


Figure 5.5. Confusion matrix obtained in the fourth round of the SVM training testing process in a specific folder.

After determining the classification model, we predicted all unlabeled passes (2,057), totaling 2,522 passes. The total sample was classified as 1,360 low difficulty passes (53.9%), 594 medium difficulty passes (23.6%) and 568 high difficulty passes (22.5%). Into each class we identify that the percentage of successful passes were 94.9, 84.0, and 49.3 for low difficulty passes, medium difficulty passes, and high difficulty passes respectively, Figure 6.

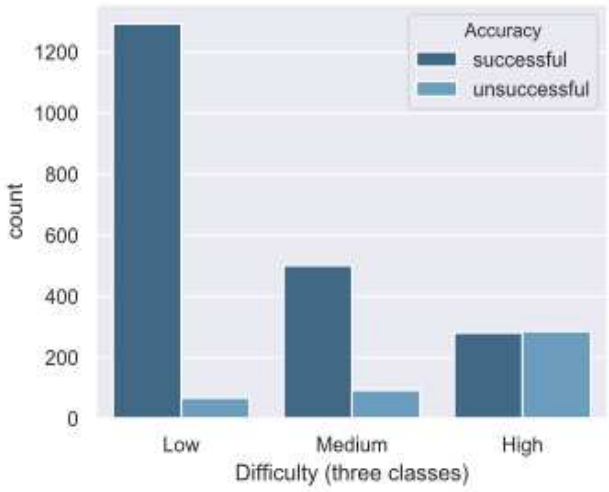


Figure 5.6. Total sample (n = 2,537) classified into three classes according to passing difficult and accuracy.

In the second step of this study, we analyze players and specific positions using a new pass classification. In an exploratory analysis, we observed the tactical demands to perform passes with different difficulty degree for each specific position. Figure 7 shows the higher proportion of medium and high difficulty passes for forwards (FW) and offensive midfields (OM) with low difficulty passes when compared with other positions.

Then we use the percentage of successful passes for each player in each class, i.e., low difficulty passes accuracy, medium difficulty passes accuracy, and high difficulty passes accuracy. We also considered the specific positions of each one. These three variables were used as an input for principal component analysis (PCA). The PCA revealed three main components (PC1, PC2, and PC3) that together explain 100% of the total sample variance. PC1, which explains 41.3% of the variance, showed a higher correlation with high difficulty passes accuracy (0.80), followed by medium difficulty passes accuracy (0.73) and low difficulty passes accuracy (0.24). PC2, which explains 33.9% of the variance, showed a greater correlation with low difficulty passes accuracy (0.93), followed by medium (0.38) and high difficulty (0.07). And PC3 explains another 24.7% of the variance showed a higher correlation with the accuracy in high difficulty passes (0.59), followed by the accuracy in medium (-0.55) and low difficulty passes (0.27). Figures 8a and 8b show the position of each player categorized by specific positions in relation to PC1, PC2, and PC3 based on their scores. When we ordered the players from the PC1 scores, we obtained the ranking of the best passing players (Table 4). In addition, we center the analysis of the scores of all players categorized by their respective positions, that is, according to the average of the specific group scores (Figure 8b). From that, we observed the position of each specific role in relation to PC1, PC2, and PC3, as well as their ranking.

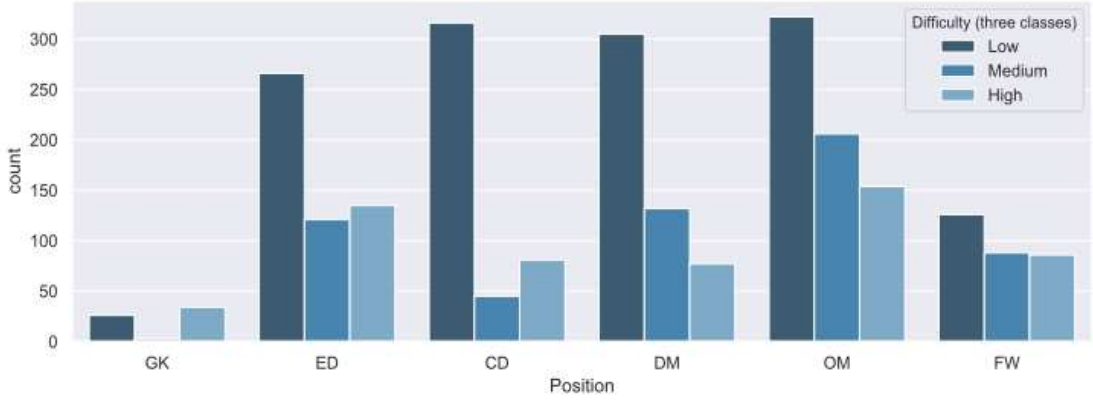


Figure 5.7. GK (goalkeeper), external defenders (ED), central defenders (CD), defensive midfield (DM), offensive midfield (OM), and forwards (FW).

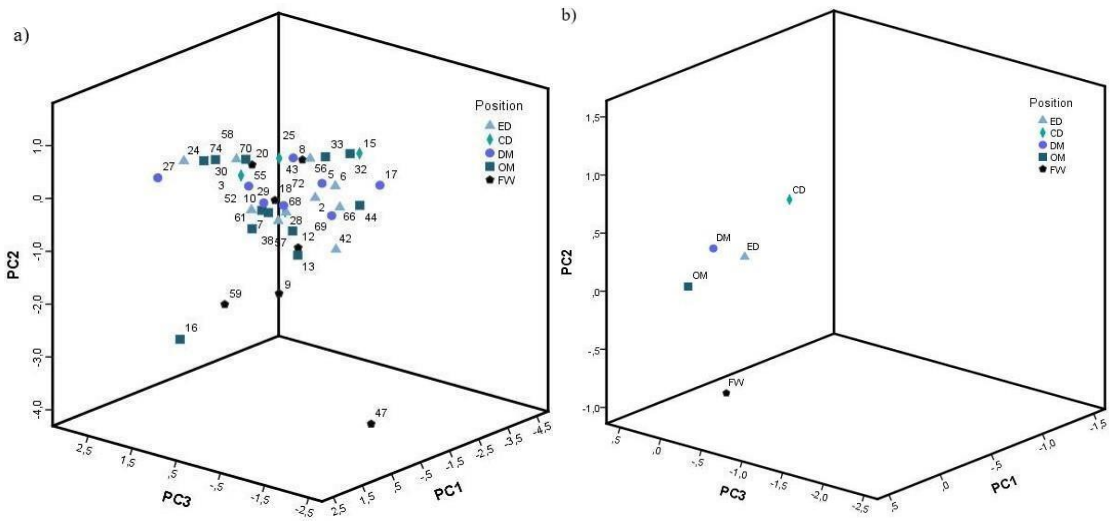


Figure 5.8. a) Three-dimension plot of the principal component analysis (PCA) of 41 players based on accuracy in low, medium and high difficulty passes, categorized by their specific positions. b) Three-dimension plot of the PCA of five positions that represent the 41 players. Abbreviations: PC1 = principal component, PC2 = principal component, PC3 = principal component. ED = external defenders, CD = central defenders, DF = defensive midfield, OM = offensive midfield, FW = forwards.

Table 5.4: Ranking of the best passing players ordered from the principal component 1.

Ranking			Accuracy (%)			Scores		
	Player	Position	Low	Medium	High	PC1	PC2	PC3
1°	16	OM	77.8	84.6	100.0	2.02	-2.32	0.99
2°	27	DM	97.8	94.7	100.0	1.82	0.55	1.62
3°	59	FW	81.8	91.7	83.3	1.63	-1.56	0.23
4°	24	ED	100.0	85.7	85.7	0.98	0.69	1.58
5°	13	OM	87.5	100.0	55.6	0.88	-0.52	-0.93
6°	18	FW	94.4	100.0	60.0	0.83	0.38	-0.38
7°	55	DM	96.4	94.4	66.7	0.79	0.49	0.24
8°	72	DM	93.8	100.0	57.1	0.76	0.30	-0.53
9°	58	OM	100.0	88.9	75.0	0.76	0.81	1.02
10°	74	OM	100.0	84.6	77.8	0.69	0.70	1.32
:								
37°	17	DM	96.2	91.7	16.7	-0.89	0.58	-1.63
38°	43	FW	100.0	75.0	37.5	-0.93	0.63	0.16
39°	20	FW	100.0	57.1	50.0	-1.17	0.18	1.46
40°	42	ED	88.9	66.7	28.6	-1.21	-0.99	-0.42
41°	68	CD	95.5	0.0	20.0	-4.05	-1.60	2.60

Abbreviations: PC1 = first principal component, PC2 = second principal component, PC3 = third principal component. ED = external defenders, CD = central defenders, DF = defensive midfield, OM = offensive midfield, FW = forwards. Low = low difficulty passing, Medium = medium difficulty passing, High = high difficulty passing.

Table 5.5: Ranking of the best passing players grouped by position ordered from the principal component 1.

Ranking	Position	Accuracy (%)			Scores		
		Low	Medium	High	PC1	PC2	PC3
1°	DM	95.9	93.0	52.8	0.31	0.44	-0.27
2°	OM	94.0	87.3	57.1	0.30	0.05	0.06
3°	FW	87.9	80.4	52.7	0.08	-0.90	-0.14
4°	ED	95.4	81.0	48.1	-0.25	0.11	0.05
5°	CD	98.3	68.7	40.4	-1.02	0.26	0.47

Abbreviations: PC1 = first principal component, PC2 = second principal component, PC3 = third principal component. ED = external defenders, CD = central defenders, DF = defensive midfield, OM = offensive midfield, FW = forwards. Low = low difficulty passing, Medium = medium difficulty passing, High = high difficulty passing.

5.4. Discussion

This aim of this study was twofold. The first one was to classify automatically the degree of passing difficulty in soccer matches using machine learning classifiers. The support vector machine (SVM), a non-linear model, proved to be the best prediction model based on machine learning techniques capable of classifying passes with different degrees of difficulty. The SVM presented a mean of balanced accuracy of 0.70, that is, it suggests that the model has a 70% chance to correctly classify a pass in professional soccer matches in low, medium and high difficulty, based on the predictor variables proposed in that study. The SVM reached a balanced accuracy of 88% in their best performance. Figure 5.9 shows three examples of passes classified by the machine learning model in low, medium, and high difficulty.

Some considerations must be made based on the results found. It was evidenced that the models with three classes had higher values of accuracy when compared with five classes. Therefore, we will focus our discussion on this scenario. In the second point, it was clear that this is an unbalanced classification problem. The sample labelled by experts was distributed with 56.6% for the low difficulty passes (class 1), 22.6% for the medium difficulty passes (class 2), and 20.9% for the high difficulty passes (class 3). For this we chose balanced accuracy as the ideal metric to compare classifiers. The balanced accuracy considers the average of the ratio between true positives plus true negatives for the total sample in each class, that is, the maximum capacity of the model to classify correctly, considering each class, thus avoiding overestimating the classifier's performance. And in the third point, the experts agreed in 80.2% of the cases when labeling the 465 events in

three classes. This percentage suggests the subjective nature of the problem, which makes it difficult to achieve very high accuracy values.

Similar research has also been using passing prediction models in professional soccer matches. The passing ability model is based on probability that each pass is successful, given information on the environment in which the pass was made and the identity of the player making the pass (Mchale, 2015). Mchale & Relton (2018) aimed to identify key players using network analysis and difficulty passes, but defined difficulty as a synonym for importance and also assumed as a criterion the probability to complete the pass. Power et al. (2017) proposed a logistic regression model to assess the risk and advantage of the pass. The risk is conditioned by the likelihood that the player will successfully make the pass given a player has possession of the ball and the advantage the likelihood that the pass made will result in a shot within the next 10 seconds. They assign higher values for passes less likely to be completed. The present study proposed a set of variables different from the others, aiming to contemplate technical and tactical attributes of the match, i.e., observing and judging the degree of difficulty of each event based on the proposed concept.

Other recent studies have also had the challenge of improving information about the passes in soccer matches, through metrics to measure effectiveness (Bransen & Haaren, 2019; Goes et al., 2018; Rein et al., 2017), probabilistic models (Spearman et al., 2017), indexes (Cintia et al., 2015), among others. In the study that most resembles ours, (Horton et al., 2014) obtained 90.0% accuracy to classify quality of the pass in good, ok and bad, using a logistic regression model. These results are higher than the present study, but the accuracy values are unbalanced. In addition, unsuccessful passes were excluded.

As a general idea, the studies start from the same principle, that is, to assign greater weight in the efficiency in performing more difficult passes, either through regression where the outputs are continuous values, or classification where the outputs are categorical. The present research brings some fundamental differences. The concept of passing difficulty was originally proposed and is essential to our problem. The focus of the experts when labeling passes was centered in the degree of technical and tactical demands that the passing player must complete the action successfully. Furthermore, in our conception, there is a difference between difficulty and quality or advantage of the passes. We focused on the difficulty because we wanted to analyze the player's ability to perform passes relativizing by the degree of difficulty.

Normally, in professional first division matches, players presented an average success rate of 84.3%, as an English Premier League games (Power et al., 2017). In our sample, the first division of Brazilian soccer, we observed a success rate of 82.3%. When we observed the percentage of successful passes in each class, we found that high difficulty passes had a success rate of 49.3% only, followed by 84.0% for medium difficulty passes and 94.9% for low difficulty passes (Figure 6). These data confirmed the need to relativize accuracy by the degree of difficulty, and provided the second part of our aim, which was to use this information to discriminate between players and positions.

In the first application from the complete sample, 2,522 passes classified within the three classes and categorized into successful and unsuccessful passes, we seek to understand the demand of each player and position, and to discriminate the best passing players. The exploratory analysis from frequency of occurrence (Figure 7) showed that the proportion of low, medium, and high difficulty passes is different between positions, where a higher proportion of high difficulty passes for forwards (FW) and offensive midfields (OM) in relation to the other positions. However, it was necessary to analyze the performance of each player and position for each class. Therefore, these three variables, accuracy in low, medium and high difficulty passes, were used as inputs for principal component analysis (PCA). The PCA revealed three principal components that together explained 100% of the variance contained in the three predictor variables. PC1, which explains most of the variance, showed a higher correlation with the variable accuracy in high and medium difficulty passes. This finding suggests that it is more important to consider the player's ability to complete high and medium difficulty passes than low difficulty passes. When we observed the players' ranking from the PC1 score, the best ranked players, players 16 and 27, showed 100.0% efficiency in high difficulty passes. On the other hand, the last ranked, player 68, although he showed 95.5% efficiency in low difficulty passes, showed only 20.0% efficiency in high difficulty passes.

When analyzing the players positions, we observe that the two main passing players based on PC1 were midfields, and the worst ranked is center defender. Table 5 showed that the highest scores from PC1 are DM, OM, FW, ED and CD, respectively. This confirms our suspicion that midfielders and forwards are better able to complete high difficulty passes, supposedly because they are technically better. This finding corroborates the study by (Bransen & Haaren, 2019), who identified the midfielders with the lowest scores in the metric (ECOM) proposed by them. ECOM measures the players'

involvement in setting up chances by valuing the effectiveness of their passes. Other studies have also applied their metrics to rank players and positions. (Mchale & Relton, 2018) used a statistical model to determine the difficulty of a pass and combined this information with results from network analysis, to identify which players are pivotal to each team. They also highlighted midfielders and forwards when compared with other positions.

The present study aimed to improve the level of information on the most frequent and determinants technical-tactical action in soccer matches. Classifying the pass in different degrees of difficulty proved to be important when analyzing the efficiency relativized by the complexity of the performed action. The player ability in performing more difficult tasks, in this case the pass, was determinant when discriminating players and positions, and it can also contribute to discriminate teams, analyze offensive sequences, and identify talents.

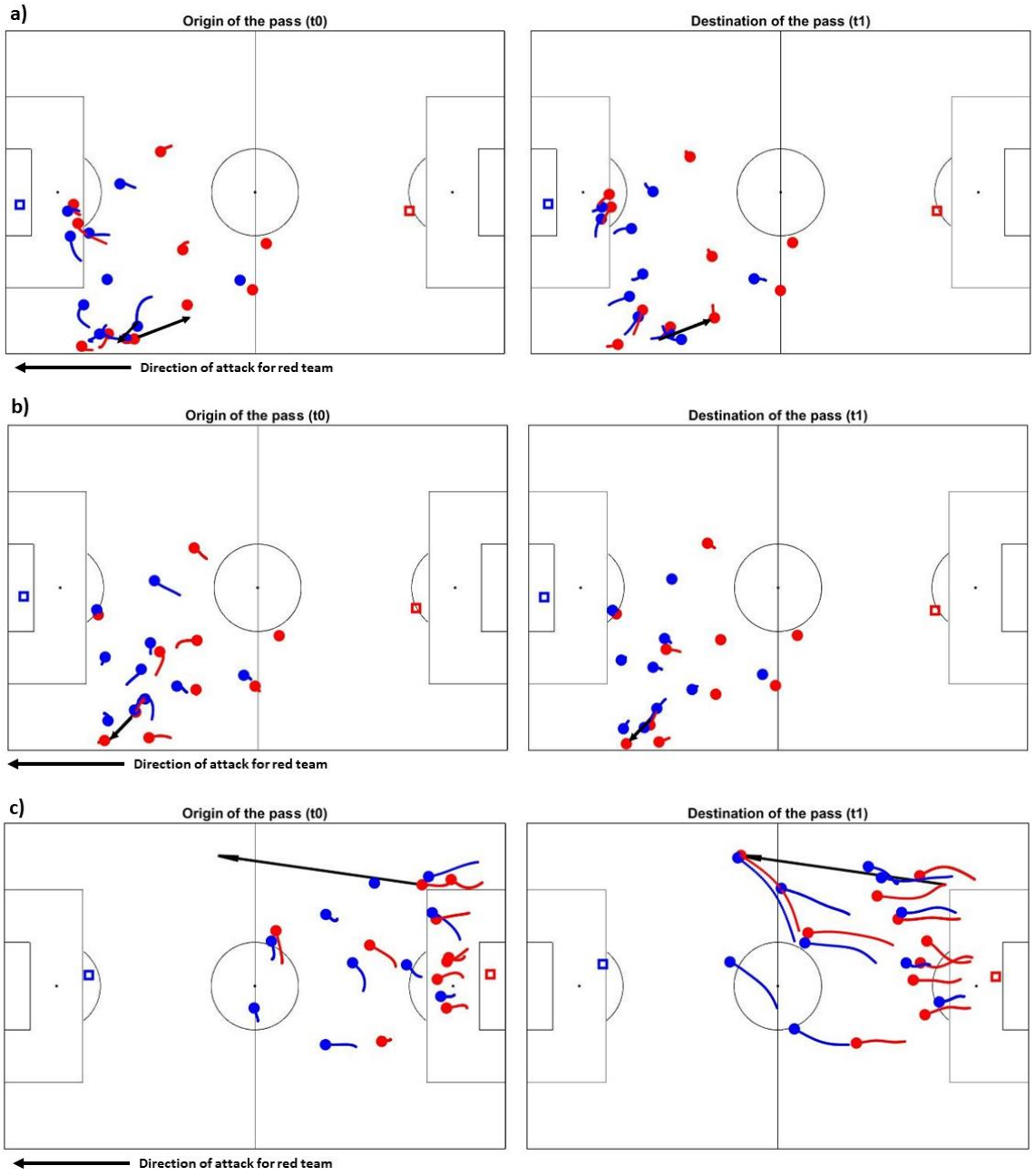


Figure 5.9. Illustration of real pass situation classified by machine learning (ML) model. Origin of the pass = at the moment of contact with the ball (t0); Destination of the pass = at the moment of reception (t1). a) Example of low difficulty pass classified by ML. b) Example of medium difficulty pass classified by ML. c) Example of high difficulty pass classified by ML. Red team attacks to the left and blue team attacks to the right.

5.5. Conclusion

In summary, the SVM classifier showed better performance among the other classifiers when classifying passes in low, medium and high difficulty based on the predictor variables. We applied the classification model, SVM, to predict a new sample of passes. Then, we identified through the principal components analysis that the efficiency in performing high and medium difficulty passes is more determinant to distinguish players and positions, that is, the best passing players were those who had the highest percentage of successful passes in high and medium difficulty, respectively. Midfielders and forwards highlighted in relation to the other positions. The proposed model, therefore, improved the level of information about passes actions, which is the most frequent and determinant for performance in soccer matches. In addition, the model can be applied in other analyzes such as offensive sequences analysis and talent identification.

6. General discussion

The general objective of the thesis was to propose a new approach to analyze the passing in soccer matches using multivariate and machine learning techniques, i.e., from the access to the spatiotemporal data, and with the multivariate and machine learning techniques available, this research seek to break with analyzes traditionally used in the scientific and practical context. For this, we analyzed the pass from two perspectives: as a sequence of passes or ball possession (BP), and the pass as an event.

The present thesis contributed to the proposition of 77 variables, originally proposed, adapted or inspired by other studies. These variables have described technical-tactical characteristics, based on the spatiotemporal data of all players on the pitch. According to Rein & Memmert (2016), the tactics in soccer describe microscopic and macroscopic organizational principles of the player. Similarly, Power et al. (2017), describes variables obtained from the spatiotemporal data such micro-level (individual feature), and high-level (contextual information at the team). We prioritize collective variables that describe the team's behavior perform passes sequences, that is, variables related to macroscopes organizational principles or high-level feature. In this case, team's contextual information appears to be most significantly described using time series analysis. When we analyze the pass as an event, we prioritize variables that describe microscopic organizational principles, or micro-level features. Therefore, the present study contributed with variables, which describe different organizational principles, but in separate studies, Figure 6.1.

The first study, *Exploring the determinants of success in different clusters of ball possession sequences in soccer*, we address to the first perspective. We analyzed the pass within BP using 41 variables predominantly collective and dynamic. In the first step, the cluster analysis identified three groups, short, medium and long duration. This was the first contribution of this thesis because we used quantitative methods based on grouping by similarity to distinguish and identify these three BP characteristics, different of most studies that propose this division subjectively (Collet, 2013; da Mota et al., 2015; Yiannakos & Armatas, 2017). In addition, in the second step we use FDA to highlight five between forty-one tactical variables, the most relevant that better describe these three clusters: coefficient of variation (CV) of the defensive team's synchronization-Y, CV defensive team´s synchronization-X, successful pass last third, CV distance between

offensive team's centroid and target, mean of the offensive team's width. The findings provided accurate tactical characterization to offensive and defensive team´s in the short, medium and long BP and suggest collective behaviors that help to maintain BP and perform passes.

The second and third study focused their analyses on concept of passing difficulty, originally proposed in this thesis: "passing difficulty refers to the degree of technical and tactical demands that the passing player has to complete the action successfully". This concept guided the proposition of 36 variables related to the pass, pressure on the passing player, pressure on the passing receiver, ball trajectory, pitch position and passing player techniques. In both studies, we used a sample with 465 passes labelled experts. The passes were classified such low difficulty (Low-DP), medium difficult (Medium-DP) and high difficulty (High-DP).

Studies 2 and 3 are complementary. In study 2, *Classification and determinants of passing difficulty in soccer: a multivariate approach,* we opted for a more interpretive multivariate statistical model. The FDA, in addition to achieving an important accuracy in classifying passes in different degrees of difficulty, 72.0%, highlighted the variables that contributed most to the model. These variables were discussed within session II and can bring some benefits to the science of sport, more especially for the match analysis process in soccer: training planning and manipulation using as reference the values of the most important variables, reports of players, teams and opponents from the most important variables of the pass in order to identify strengths, weaknesses and characteristics when executing passes. Besides that, the discriminant function propose can be used by coaches in a practical context to analyze passing performance of their players and teams.

The study three, into the session III, *Who are the best passing players in professional soccer? Machine learning approach classifies passes with different levels of difficulty and discriminate the best passing players*, aimed to improve the prediction by classifying passes using machine learning algorithms, and to apply the model with the best performance to discriminate players and positions. The support vector machine (SVM), a non-linear model, proved to be the best prediction model based on machine learning techniques, reaching a balanced accuracy of 88% in their best performance. After, we applied the classification model, SVM, to predict a new sample of passes. Classifying passes based on the degree of difficulty allows some inferences in players and

teams. We used success rate in low, medium and high difficulty passes, as inputs for principal component analysis (PCA). The PC1, which explains most of the variance, showed a higher correlation with the variable success rate in high and medium difficulty passes, suggesting that it is more important to consider the player's ability to complete high and medium difficulty passes than low difficulty passes. In addition, when analyzing the players positions, we observe that the two main passing players based on PC1 were midfields, and the worst ranked is center defender. The player ability in performing more difficult tasks, in this case the pass, was determinant when discriminating players and positions, and it can also contribute to discriminate teams, analyze offensive sequences and identify talents. The model propose can be used by coaches in a practical context to analyze passing performance of their players and teams.

This research proposes a new approach to analyze the pass in soccer matches. This approach can be adapted to analyze other technical actions such as dribbling, shots, tackles. Historically, technical actions are analyzed based on frequency and success rate. This research proposed 77 variables with collective characteristics (macroscopic) and individual or group (microscopic) that quantitatively describes the players and team actions. The multivariate and machine learning techniques used contributed to identify patterns, prediction and highlight variables. By analyzing the degree of difficulty of the technical-tactical actions, we contribute to the construction of knowledge that can help to quantify the match complexity. This may be a new research trend in match analysis and contribute to areas of sports science.

The main limitation of this research was the relatively small sample. Although the number of events, such as possession of ball in study 1, and passes in studies 2 and 3, were sufficient for the research problem, the reduced number of matches limited some inferences with teams and players.

In addition to the contributions described in the thesis, this research leaves other possibilities open for the research in sports in general. In study 1, we compared pass sequences according to the time and number of passes. Future analyses could use similar approach to identify behaviours that explain the team's ability in offensive sequences that end in shots and goals, considering it is the main objective of the match. In studies 2 and 3, the pass can be labeled using other concepts. For example, the tactical importance, that is, the tactical advantage that the pass provided. There are studies in this sense, but with the same limitations described in this thesis about the analysis of passing difficulty. Therefore, the approach used in this study can be used to analyze the tactical importance

of the passes. Besides that, the highlighted variables can be used to identify weaknesses and strengths in players and teams. We can use the most important variables as input to a model to identify what determine the player perform successful and unsuccessful passes. The proposed model can be adapted to other contexts such as female soccer, youth soccer categories. In addition, the idea of classifying the difficulty of passes action, can be applied to other actions in the soccer matches.

Finally, future studies may propose an integrated analysis of the pass, using the variables associated with difficulty, that describe microscopic organizational principles, with collective variables represent to macroscopes organizational principles.

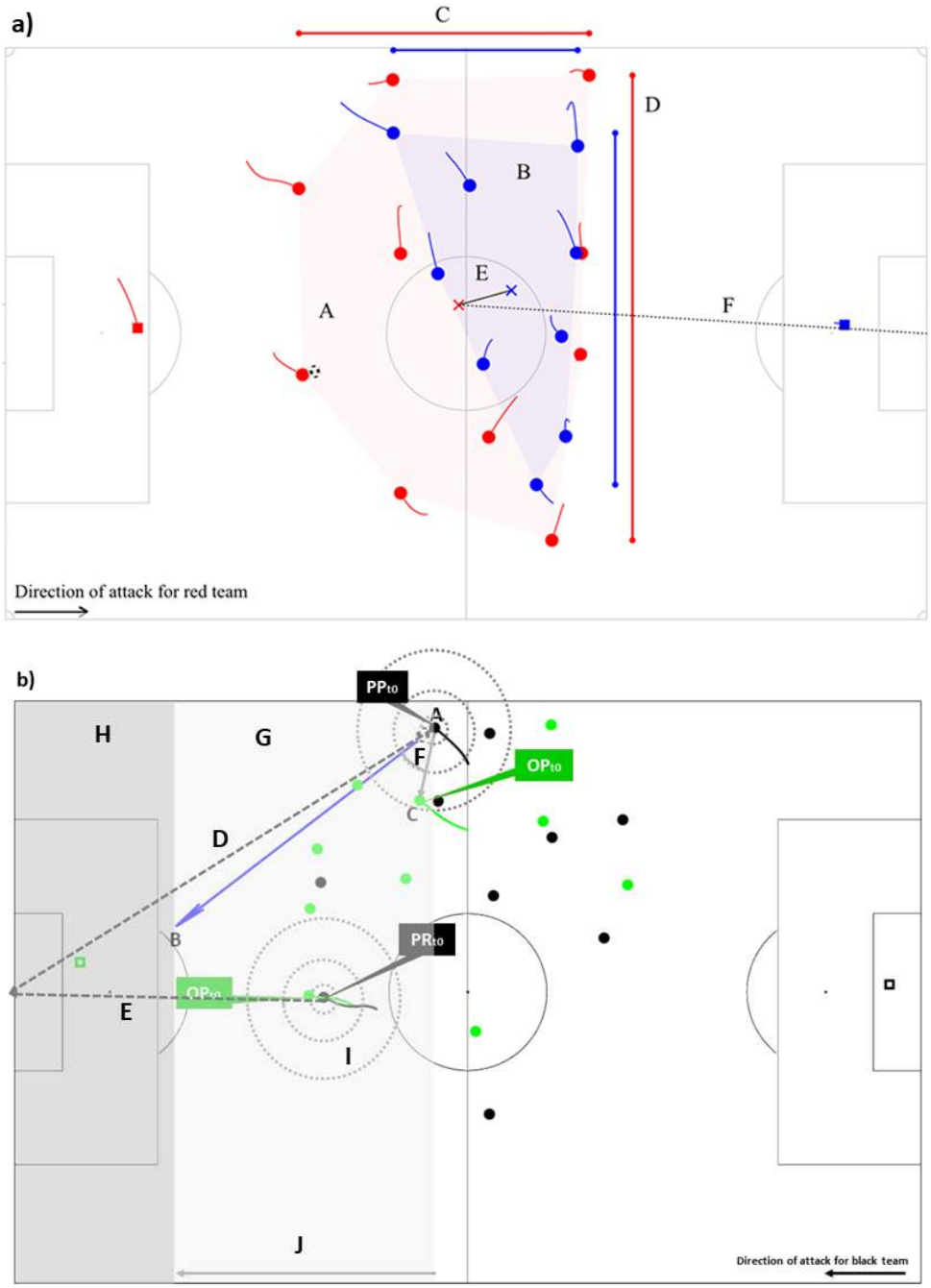


Figure 6.1. Comparative representation. a) Variables that describe the team's behavior perform passes during ball possession. Abbreviations: A = Effective playing space (red team); B= Effective playing space (blue team); C = length (red team); D = width (red team); E = distance between team centroids; F = distance between centroid and target (red team). b) Variables that describe the passing difficulty at the moment of the pass (t0). Abbreviations: PP_{t0} = passing player at t0; PR_{t0} = receiver at t0; OP_{t0} = nearest opponent to the passing player and receiver at t0; $(\overrightarrow{AB})$ = passing distance; $(\overrightarrow{AC})$ distance between passing player and his nearest opponent at t0; D = distance between passing player and target of opponent at t0; E = distance between passing receiver and target of opp. at t0; F = opponent angle; G = number of outplayed opponent (into light gray shaded area); H = opponent between PRt1 and target (into dark gray shaded area); I = number of opponents within the 1m, 2m, 5m and 10m radius to passing receiver at t1; J = Ball progression.

7. General conclusion

The present thesis aimed to break with analyzes traditionally used in the scientific and practical context and proposed a new approach for analyzing the passing in soccer matches from two perspectives, as a passes sequence and the pass as an event. A total of 77 variables was originally proposed, adapted or inspired by other studies. The analysis of passes sequences revealed the most determinant variables to discriminate short, medium and long ball possessions, suggesting collective behaviors that help to maintain BP and perform passes. Then we proposed the passing difficulty concept. Using this concept, we build two models to classify degree of difficulty passes automatically in soccer matches. The first one, using Fisher Discriminant Analysis, we highlighted the most determinant variables to discriminate low, medium and high difficulty passes. The second model using machine learning (ML) classifiers, we improve the predictive power. The Support Vector Machine (SVM) was the ML model with the best performance, achieved 88% accuracy when classifying passes in low, medium and high difficulty.

As practical applications, the passing difficulty classification models can be used by coaches in practical context. Classifying passes based on their degree of difficulty allow to analyze the ability of players and teams to execute a high-level difficulty pass. This ability has been shown to be the most relevant to rank passing players. In addition, the highlighted variables using discriminant analysis can be used to analyze individual e collective behaviors in order to improve performance in performing passes, considering that the passes are the important determinants of success in soccer matches.

Future studies may propose an integrated analysis of the pass, using the variables associated with difficulty, with collective variables, that could describe complementarily the individual and collective ability to perform passes. Besides that, to explore concepts other concepts such passing importance, to identify weaknesses and strengths in players and teams when performing passes, and adapt the approach this research to other contexts such as female soccer and youth soccer categories.

8. References

Aguiar, M., Gonçalves, B., Botelho, G., Duarte, R., & Sampaio, J. (2017). Regularity of interpersonal positioning discriminates short and long sequences of play in small-sided soccer games. *Science and Medicine in Football*, *1*(3), 258–264. https://doi.org/10.1080/24733938.2017.1353220

Baron, R., Tschan, H., Montero, F. J. C., & Bachl, N. (2006). Performance Characteristics According to Playing Position in Elite Soccer. https://doi.org/10.1055/s-2006-924294

Barros, R., Misuta, M., Menezes, R., Figueroa, P., Moura, F., Cunha, S., … Leite, N. (2007). Analysis of the distances covered by first division Brazilian soccer players obtained with an automatic tracking method. *Journal of Sports Science and Medicine*, *6*, 233–242.

Bonaccorso, G. (2017). *Machine Learning Algorithms*. Birmingham.

Bradley, P. S., Carling, C., Archer, D., Roberts, J., Dodds, A., Di Mascio, M., … Krustrup, P. (2011). The effect of playing formation on high-intensity running and technical profiles in English FA Premier League soccer matches. *Journal of sports sciences*, *29*(8), 821–830. https://doi.org/10.1080/02640414.2011.561868

Bradley, P. S., Carling, C., Gomez, A., Hood, P., Barnes, C., Ade, J., … Mohr, M. (2013). Human Movement Science Match performance and physical capacity of players in the top three competitive standards of English professional soccer. *Human Movement Science*, *32*(4), 808–821. https://doi.org/10.1016/j.humov.2013.06.002

Bransen, L., & Haaren, J. Van. (2019). Measuring soccer players ' contributions to chance creation by valuing their passes, *15*(2), 97–116.

Buchheit, M. (2014). Integrating different tracking systems in football : multiple camera semi-automatic system , local position measurement and GPS technologies, (February 2016). https://doi.org/10.1080/02640414.2014.942687

Bush, M., Barnes, C., Archer, D. T., Hogg, B., & Bradley, P. S. (2015). Evolution of match performance parameters for various playing positions in the English Premier League. *Human Movement Science*, *39*, 1–11. https://doi.org/10.1016/j.humov.2014.10.003

Carling, C., Bloomfield, J., Nelsen, L., & Reilly, T. (2008). The Role of

Motion Analysis in Elite Soccer Work Rate Data, *38*(10), 839–862.

Carling, C., Reilly, T., & Williams, A. M. (2009). *Performance Assessment for fiels sports* (1º ed). New York.

Castillo, D., Raya-gonzález, J., Clemente, F. M., & Yanci, J. (2019). The influence of youth soccer players 'sprint performance on the different sided games' external load using GPS devices. *Research in Sports Medicine*, *28*(02), 1–12. https://doi.org/10.1080/15438627.2019.1643726

Chassy, P. (2013). Team Play in Football : How Science Supports F . C . Barcelona ' s Training Strategy, *4*(9), 7–12.

Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., & Malvaldi, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. In *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*. Paris, France. https://doi.org/10.1109/DSAA.2015.7344823

Clemente, F. M., Cauceiro, M. S., Martins, F. M., Ivanova, M. O., & Mendes, R. (2013). Activity Profiles Soccer During the 2010 World Cup. *Journal of Human Kinetics.*, *38*, 201–211.

Cohen, J. (1968). Psychological Bulletin, *70*(4), 213–220.

Collet, C. (2013). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010. *Journal of Sports Sciences*, *31*(2), 123–136. https://doi.org/10.1080/02640414.2012.727455

Couceiro, M. S., Martins, F. M. L., Figueiredo, A. J., Mendes, R. S., & Clemente, F. M. (2014). Soccer team's tactical behaviour: Measuring territorial domain. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, *229*(1), 58–66. https://doi.org/10.1177/1754337114547064

Coutinho, D., Gonçalves, B., Santos, S., Travassos, B., Del, P., Sampaio, J., … Travassos, B. (2019). Effects of the pitch configuration design on players ' physical performance and movement behaviour during soccer small-sided games. *Research in Sports Medicine*, *27*(3), 298–313. https://doi.org/10.1080/15438627.2018.1544133

Cunha, S. A., Moura, F. A., Santiago, P. R. P., Castellani, R. M., & Barbieri, F. A. (2011). *Futebol : aspectos multidisciplinares para o ensino e treinamento* (Edicação 1). Rio de Janeiro: Editora Guanabara

Koogan Ltda.

da Mota, G. R., Thiengo, C. R., Gimenes, S. V., & Bradley, P. S. (2015). The effects of ball possession status on physical and technical indicators during the 2014 FIFA World Cup Finals. *Journal of Sports Sciences*, *0414*(January 2016), 1–8. https://doi.org/10.1080/02640414.2015.1114660

Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal ofMachine Learning Research*, *7*, 1–30.

Dick, U., & Brefeld, U. (2019). Learning to Rate Player Positioning in Soccer, *7*(1), 71–82. https://doi.org/10.1089/big.2018.0054

Fernandez-navarro, J., Fradua, L., Zubillaga, A., Ford, P. R., Mcrobert, A. P., Fradua, L., … Fernandez-navarro, J. (2016). Attacking and defensive styles of play in soccer : analysis of Spanish and English elite teams elite teams. *Journal of Sports Sciences*, *0414*(April). https://doi.org/10.1080/02640414.2016.1169309

Fernandez, J., & Bornn, L. (2018). Wide Open Spaces : A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conferece*, (March), 1–19.

Figueroa, P. J., Leite, N. J., & Barros, R. M. L. (2006). Background recovering in outdoor image sequences : An example of soccer players segmentation. *Image and Vision Computing*, *24*, 363–374. https://doi.org/10.1016/j.imavis.2005.12.012

Folgado, H., Duarte, R., Fernandes, O., & Sampaio, J. (2014). Competing with lower level opponents decreases intra-team movement synchronization and time-motion demands during pre-season soccer matches. *PLoS ONE*, *9*(5). https://doi.org/10.1371/journal.pone.0097145

Folgado, H., Duarte, R., Marques, P., Gonçalves, B., & Sampaio, J. (2018). Exploring how movement synchronization is related to match outcome in elite professional football. *Science and Medicine in Football*, *2*(2), 101–107. https://doi.org/10.1080/24733938.2018.1431399

Folgado, H., Gonçalves, B., Sampaio, J., Folgado, H., & Gonçalves, B. (2017). Positional synchronization affects physical and physiological responses to preseason in professional football (soccer) responses to preseason in professional football (soccer). *Research in Sports Medicine*, *26*(1), 51–63.

https://doi.org/10.1080/15438627.2017.1393754

Folgado, H., Lemmink, K. A. P. M., Frencken, W., & Sampaio, J. (2014). Length, width and centroid distance as measures of teams tactical performance in youth football. *European Journal of Sport Science*, *14*(sup1), S487–S492. https://doi.org/10.1080/17461391.2012.730060

Fonseca, S., Milho, J., Passos, P., Araújo, D., & Davids, K. (2012). Approximate entropy normalized measures for analyzing social neurobiological systems. *Journal of Motor Behavior*, *44*(3), 179–183. https://doi.org/10.1080/00222895.2012.668233

Fujimura, A., & Sugihara, K. (2005). Geometric Analysis and Quantitative Evaluation of Sport, *36*(6), 49–58. https://doi.org/10.1002/scj.20254

Goes, F., Kempe, M., Lemmink, K., Goes, F., Kempe, M., & Lemmink, K. (2019). Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data Predicting match outcome in professional Dutch soccer using tactical performance metrics computed from position tracking data.

Goes, F., Kempe, M., Meerhoff, M., & Lemmink, K. (2018). Not Every Pass Can Be an Assist : A Data-Driven Model to Measure Pass Effectiveness in Professional Soccer Matches. *Big Data*, *6*(4), 1–28. https://doi.org/10.1089/big.2018.0067

Goes, F. R., Meerhoff, L. A., O, B. M. J., Rodrigues, D. M., Moura, F. A., Elferink-Gemser, M. T., … Lemmink, K. A. P. M. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer : A systematic review. *European Journal of Sport Science*, *0*(0), 1–16. https://doi.org/10.1080/17461391.2020.1747552

Gonçalves, B., Coutinho, D., Santos, S., Lago-Penas, C., Jiménez, S., & Sampaio, J. (2017). Exploring team passing networks and player movement dynamics in youth association football. *PLoS ONE*, *12*(1), 1–13. https://doi.org/10.1371/journal.pone.0171156

Gonçalves, B., Marcelino, R., Torres-Ronda, L., Torrents, C., & Sampaio, J. (2016). Effects of emphasising opposition and cooperation on collective movement behaviour during football small-sided games. *Journal of Sports Sciences*, *34*(14), 1346–1354. https://doi.org/10.1080/02640414.2016.1143111

Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). Searching for a Unique Style in Soccer. *arXiv*, 5–8. Recuperado de

http://arxiv.org/abs/1409.0308

Gyarmati, L., & Stanojevic, R. (2016). QPass : a Merit-based Evaluation of Soccer Passes Field value. *arXiv.org*. Recuperado de https://arxiv.org/abs/1608.03532

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction* (2º ed). Stanford.

Hewitt, A., Greenham, G., & Norton, K. (2016). Game style in soccer: what is it and can we quantify it? *International Journal of Performance Analysis in Sport*, (April), 355–372.

Hopkins, W. G., Marshall, S. W., Batterham, A. M., & Hanin, J. (2009). Progressive Statistics for Studies in Sports Medicine and Exercise Science. *MEDICINE & SCIENCE IN SPORTS & EXERCISE*, *41*(1), 3–12. https://doi.org/10.1249/MSS.0b013e31818cb278

Horton, M., & Gudmundsson, J. ([s.d.]). Spatiotemporal Data.

Horton, M., Gudmundsson, J., Chawla, S., & Estephan, J. (2014). Classification of Passes in Football Matches using Spatiotemporal Data. *ACM Transactions on Spatial Algorithms and Systems*, *3*(2). https://doi.org/10.1145/3105576

Hughes, M. D., & Bartlett, R. M. (2010). The use of performance indicators in performance analysis The use of performance indicators in performance analysis. *Journal of Sports Sciences*, *0414*(20), 739–754. https://doi.org/10.1080/026404102320675602

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of sports sciences*, *23*(5), 509–514. https://doi.org/10.1080/02640410410001716779

Kite, C. S., & Nevill, A. (2017). The Predictors and Determinants of Inter-Seasonal Success in a Professional Soccer Team by, *58*(September), 157–167. https://doi.org/10.1515/hukin-2017-0084

Kumar, G. (2014). *Machine Learning for Soccer Analytics*. https://doi.org/10.13140/RG.2.1.4628.3761

Lago, C., & Martín, R. (2007). Determinants of possession of the ball in soccer. *Journal of sports sciences*, *25*(9), 969–974. https://doi.org/10.1080/02640410600944626

Link, D., Hoernig, M., Nassis, G., Laughlin, M., & Witt, J. de. (2017). Individual ball possession in soccer. *Plos One*, *12*(7), e0179953. https://doi.org/10.1371/journal.pone.0179953

Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS ONE*, *11*(12), 1–16. https://doi.org/10.1371/journal.pone.0168768

Lorenzo-martínez, M., Rey, E., & Padrón-cabo, A. (2019). The effect of age on between-match physical performance variability in professional soccer players variability in professional soccer players. *Research in Sports Medicine*, *28*(03), 351–359. https://doi.org/10.1080/15438627.2019.1680985

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods : Concerns and ways forward. *PLoS ONE*, (Ml), 1–26. https://doi.org/10.1371/journal.pone.0194889

Mchale, I. (2015). Beyond completion rate : evaluating the passing.

Mchale, I. G., & Relton, S. D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, *268*(1), 339–347. https://doi.org/10.1016/j.ejor.2018.01.018

Mclaren, S. J., Hurst, C., Spears, I. R., & Weston, M. (2017). The Relationships Between Internal and External Measures of Training Load and Intensity in Team Sports : A Meta-Analysis. *Sports Medicine*, (Ci). https://doi.org/10.1007/s40279-017-0830-z

Memmert, D., Lemmink, K. A. P. M., & Sampaio, J. (2016a). Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sports Medicine*, 1–10. https://doi.org/10.1007/s40279-016-0562-5

Memmert, D., Lemmink, K. A. P. M., & Sampaio, J. (2016b). Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sports Medicine*, 1–10. https://doi.org/10.1007/s40279-016-0562-5

Mining, D. ([s.d.]). Springer Series in Statistics The Elements of.

Mitschke, C., & Milani, T. L. (2014). Soccer : Detailed Analysis of Played Passes in the UEFA Euro 2012, *9*(5), 1019–1032.

Montoliu, R., Martín-félez, R., & Torres-sospedra, J. (2015). Human

Movement Science Team activity recognition in Association Football using a Bag-of-Words-based method. *Human Movement Science*, *41*, 165–178. https://doi.org/10.1016/j.humov.2015.03.007

Moura, F., Martins, L. E., Anido, R., Barros, R., & Cunha, S. (2012). Quantitative analysis of Brazilian football players' organisation on the pitch. *Sports Biomechanics*, *11*(1), 85–96.

Moura, F., Martins, L. E., Anido, R., Ruffino, P. R., Barros, R., & Cunha, S. (2013). A spectral analysis of team dynamics and tactics in Brazilian football. *Journal of Sports Sciences*, *31*(14), 37–41.

Osgnach, C., Poser, S., Bernardini, R., Rinaldo, R., & Di Prampero, P. (2009). Energy Cost and Metabolic Power in Elite Soccer: A New Match Analysis Approach. *Medicine and Science in Sports and Exercise*, 170–178.

Ouellette, J. (2004). Principles of Play for Soccer. *Strategies*, *17*(October 2014), 3. https://doi.org/10.1080/08924562.2004.10591082

P. D. Jones, N. J. and S. D. M. D. (2004). Possession as a performance indicator in soccer. *Department of Sports Science*, (August), 98–102. https://doi.org/10.1017/CBO9781107415324.004

Paixão, P., Sampaio, J., Almeida, C. H., & Duarte, R. (2015). How does match status affects the passing sequences of top-level European soccer teams ? *International Journal of Performance Analysis in Sport*, *15*(1), 229–240.

Pascual, F., Leite, N., & Barros, R. (2002). A flexible software for tracking of markers used in human motion analysis. *Computer Methods and Programs in Biomedicine.*, *72*, 155–165.

Pedhazur, E. J., & Manning, S. (1997). *Multiple Regression in Behavioral Research* (Trird). Florida: Christopher P. Klein.

Pedregosa, F., Weiss, R., & Brucher, M. (2011). Scikit-learn : Machine Learning in Python. *Journal ofMachine Learning Research*, *12*, 2825–2830.

Peña, J. L., & Navarro, R. S. (2015). Who can replace Xavi? A passing motif analysis of football players, 9. Recuperado de http://arxiv.org/abs/1506.07768

Pereira, T. J. C., van Emmerik, R. E. A., Misuta, M. S., Barros, R. M. L., & Moura, F. A. (2017). Interpersonal coordination analysis of tennis

players from different levels during official matches. *Journal of Biomechanics*, *67*, 106–113. https://doi.org/10.1016/j.jbiomech.2017.11.036

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, *88*(6), 2297–2301. https://doi.org/10.1073/pnas.88.6.2297

Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). Not All Passes Are Created Equal: Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1605–1613. https://doi.org/10.1145/3097983.3098051

Reep, C., & Benajmin, B. (1968). Skill and Chance in Association Football. *Journal of the Royal Statistical Society*, *131*(4), 581–585. Recuperado de http://www.jstor.org/stable/2343726 .

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springer Plus*, *5*, 1–13.

Rein, R., Raabe, D., & Memmert, D. (2017). Human Movement Science " Which pass is better ? " Novel approaches to assess passing e ff ectiveness in elite soccer. *Human Movement Science*, *55*(August), 172–181. https://doi.org/10.1016/j.humov.2017.07.010

Riahi, Y., & Riahi, S. (2018). Big Data and Big Data Analytics : Concepts , Types and Technologies, *5*(9), 524–528.

Rico-gonzález, M., Arcos, A. L., Nakamura, F. Y., Arruda, F., & Pino-ortega, J. (2019). The use of technology and sampling frequency to measure variables of tactical positioning in team sports : a systematic review. *Research in Sports Medicine*, *00*(00), 1–14. https://doi.org/10.1080/15438627.2019.1660879

Sampaio, J., & Ma????s, V. (2012). Measuring tactical behaviour in football. *International Journal of Sports Medicine*, *33*(5), 395–401. https://doi.org/10.1055/s-0031-1301320

Sarnento, H., Marcelino, R., Anguera, M. T., Campaniço, J., Matos, N., & Leitão, J. C. (2014). Match analysis in football : a systematic review. *Journal of Sports Sciences*, *32*(20), 1831–1843. https://doi.org/10.1080/02640414.2014.898852

Sikka, R. S., Baer, M., Raja, A., Stuart, M., & Tompkins, M. (2019).

Analytics in Sports Medicine Implications and Responsibilities That Accompany the Era of Big Data. *The Journal of Bone and Joint Surgery*, *101*(3), 276–283.

Sparrow, W. A., Donovan, E., Van Emmerik, R., & Barry, E. B. (1987). Using relative motion plots to measure changes in intra-limb and inter-limb coordination. *Journal of Motor Behavior*, *19*(1), 115–129. https://doi.org/10.1080/00222895.1987.10735403

Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-Based Modeling of Pass Probabilities in Soccer. In *Sports Analytics Conference* (p. 1–14).

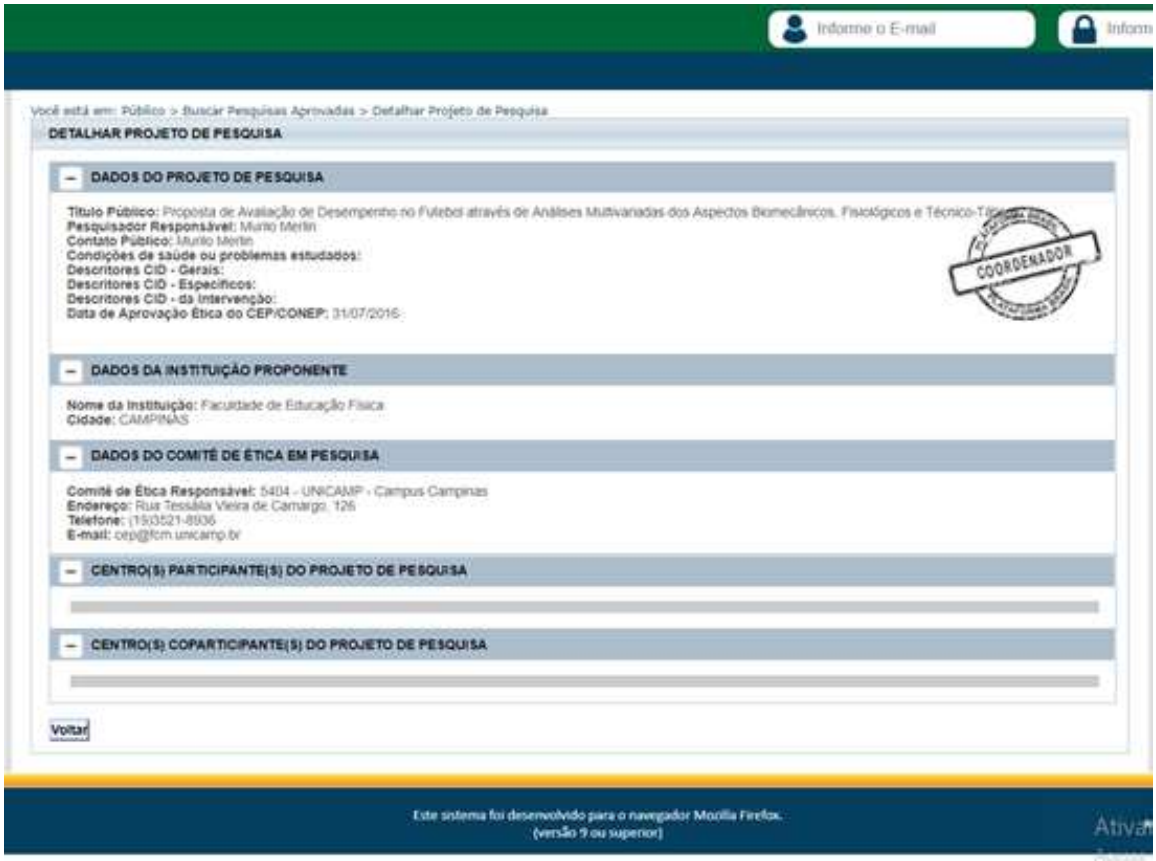
Steiner, S. (2018). Passing Decisions in Football: Introducing an Empirical Approach to Estimating the Effects of Perceptual Information and Associative Knowledge. *Frontiers in Psychology*, *9*(March), 1–11. https://doi.org/10.3389/fpsyg.2018.00361

Wallace, J. L., & Norton, K. I. (2014). Evolution of World Cup soccer final games 1966-2010: Game structure, speed and play patterns. *Journal of Science and Medicine in Sport*, *17*(2), 223–228. https://doi.org/10.1016/j.jsams.2013.03.016

Whitaker, G. A., Silva, R., & Edwards, D. (2018). Visualizing a Team ' s Goal Chances in Soccer from Attacking Events : A Bayesian Inference Approach. *Big Data*, *6*(4), 271–290. https://doi.org/10.1089/big.2018.0071

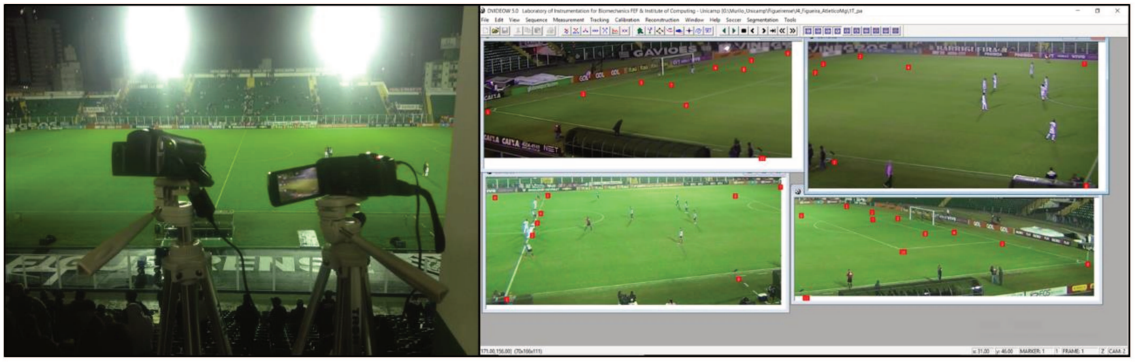
Yiannakos, A., & Armatas, V. (2017). Evaluation of the goal scoring patterns in European Championship in Portugal 2004 . *International Journal of Performance Analysis in Sport*, *6*(1), 178–188. https://doi.org/10.1080/24748668.2006.11868366

Appendix A. The Ethics Committee of the Campinas State University.

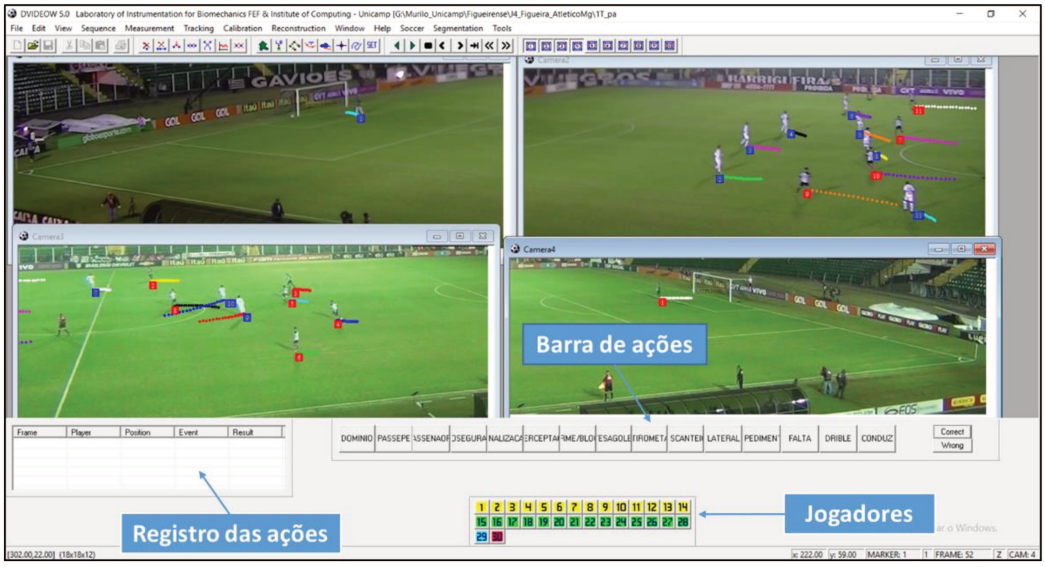


Appendix B. Images of data collection: a) camera positioning and calibration points. b) Interface DVideo software performing a notational analysis.

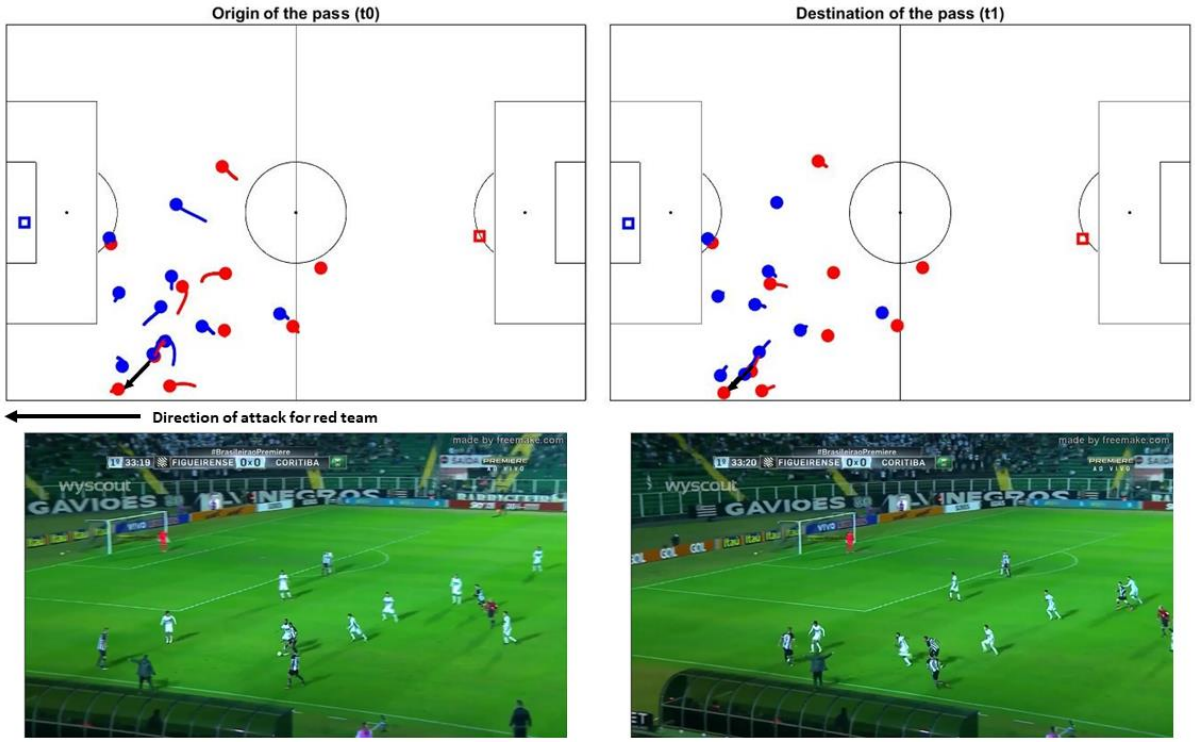
a)



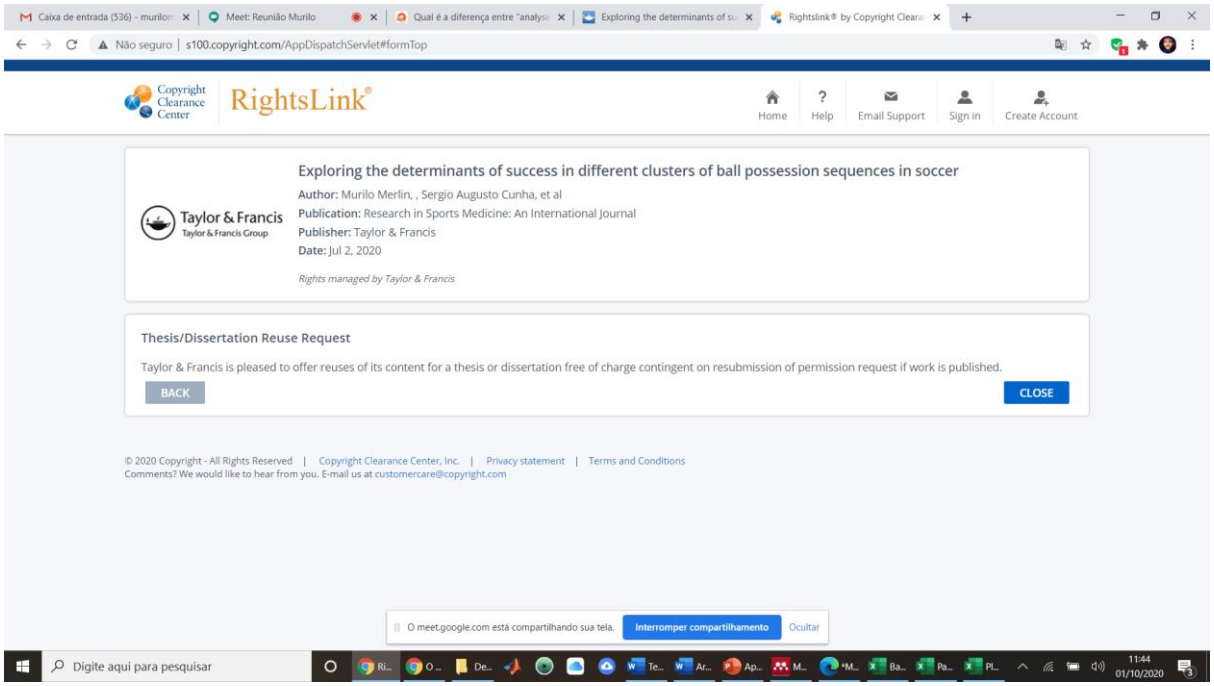
b)



Appendix C. Comparison between real (TV camera) and 2D image for a given pass at two different times, origin of the pass (t0) and destination of the pass (t1).



Appendix D. Publisher authorization.



To link to this article: https://doi.org/10.1080/15438627.2020.1716228