Predicting Football Match Result Using Fusion-based Classification Model

Mr. Chananyu Pipatchatchawal

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science and Information Technology
Department of Mathematics and Computer Science
FACULTY OF SCIENCE
Chulalongkorn University
Academic Year 2020

การทำนายผลการแข่งขันฟุตบอลโดยใช้ตัวแบบการจำแนกประเภทบนพื้นฐานการรวมตัว

นายชนัญญู พิพัฒน์ชัชวาล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	Predicting Football Match Result Using Fusion-based Classification Model
By	Mr. Chananyu Pipatchatchawal
Field of Study	Computer Science and Information Technology
Thesis Advisor	Associate Professor SUPHAKANT PHIMOLTARES, Ph.D.

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

.... Dean of the FACULTY OF SCIENCE
(Professor POLKIT SANGVANICH, Ph.D.)

THESIS COMMITTEE

.... Chairman
(Professor CHIDCHANOK LURSINSAP, Ph.D.)

.... Thesis Advisor
(Associate Professor SUPHAKANT PHIMOLTARES, Ph.D.)

.... External Examiner
(Prem Junsawang, Ph.D.)

ชนัญญู พิพัฒน์ชัชวาล : การทำนายผลการแข่งขันฟุตบอลโดยใช้ตัวแบบการจำแนกประเภทบนพื้นฐานการรวมตัว. (Predicting Football Match Result Using Fusion-based Classification Model) อ.ที่ปรึกษาหลัก : รศ. ดร.ศุภกานต์ พิมลธเรศ

ในช่วงหลายทศวรรษที่ผ่านมา ผู้เชี่ยวชาญและนักวิจัยหลากหลายคนทั่วโลกพยายามแสวงหาวิธีพยากรณ์ผลการแข่งขันฟุตบอล แม้ว่าจะมีตัวแบบการทำนายผลหลายชนิดได้สร้างขึ้นมาเพื่อจุดประสงค์นี้ ตัวแบบส่วนใหญ่ยังคงพึ่งการรวมตัวเลขทางสถิติระหว่างเกม เช่น จำนวนการผ่านบอลสำเร็จในหนึ่งเกม ข้อมูลจำพวกนี้ส่งผลกระทบทางบวกต่อการทำนายผลอย่างมาก แต่ก็ไม่เป็นที่ปรารถนาเนื่องจากมีความจำเป็นต้องให้การแข่งขันนั้นเสร็จสิ้นก่อน ดังนั้นวิทยานิพนธ์ฉบับนี้มีจุดประสงค์เพื่อเสนอตัวแบบที่แม่นยำมากกว่าโดยไม่พึ่งตัวเลขระหว่างเกม ตัวแบบการจำแนกบนพื้นฐานการรวมตัวสองชนิดถูกเสนอในการศึกษานี้ ซึ่งประกอบด้วยตัวแบบชนิดลำดับขั้นและตัวแบบชนิดรวมตัว ค่าประเมินของผู้เล่นและทีมจากวิดีโอเกมถูกนำมาใช้ร่วมกันเพื่อช่วยในการทำนายผลการแข่งขัน การทดลองถูกออกแบบมาเพื่อเปรียบเทียบตัวแบบที่เสนอกับตัวแบบดั้งเดิมหกชนิดโดยการวัดความแม่นยำด้วยชุดข้อมูลที่รวบรวมจากการแข่งขันฟุตบอลพรีเมียร์ลีกอังกฤษจากฤดูกาล 2010/2011 ถึง 2015/2016 ค่าความแม่นยำที่ได้จากทั้งสองตัวแบบอยู่ที่ 56.533% และ 56.800% โดยการใช้คุณลักษณะทั้งหมด 36 อย่าง ซึ่งดีกว่าตัวแบบที่เปรียบเทียบในสภาพแวดล้อมเดียวกัน

สาขาวิชา	วิทยาการคอมพิวเตอร์และ เทคโนโลยีสารสนเทศ	ลายมือชื่อนิสิต ....
ปีการศึกษา	2563	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6278006723 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

KEYWORD: prediction football match result, fusion-based classification, hierarchical model, ensemble model

Chananyu Pipatchatchawal : Predicting Football Match Result Using Fusion-based Classification Model. Advisor: Assoc. Prof. SUPHAKANT PHIMOLTARES, Ph.D.

Over the last decades, various football experts and researchers around the world seek to forecast football match result. Although there are multiple types of prediction models constructed for this purpose, most of them still depends on integrating in-game statistical numbers, such as number of successful passes in one game. This kind of information has huge positive impact on predicting outcome but is not desired as it requires the match to finish first. Thus, this thesis aims to propose more accurate models, which are not relied on in-game numbers. Two forms of fusion-based classification models are proposed in this study, including hierarchical model and ensemble model. Player and team ratings from video games are incorporated to assist in match prediction. Experiments are designed to compare the proposed models with six traditional classification models, by measuring accuracy with data sets collected from English Premiere League (EPL) competition, from season 2010/2011 to 2015/2016. Accurate rates achieved by both models are at 56.533% and 56.800% by using a total of 36 features, which are all superior to comparative models with similar environment.

Field of Study:	Computer Science and Information Technology	Student's Signature ....
Academic Year:	2020	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to all people who make this thesis possible.

First, this thesis could not be completed without suggestions and time devoted by my advisor, Associate Professor Dr. Suphakant Phimoltares. I am very thankful to be under his guidance during the study.

I would also appreciate all of my thesis committee: Professor Chidchanok Lursinsap and Dr. Prem Junsawang for their insightful comments and time.

Furthermore, I would like to thank my mother and friends for their support, encouragement, and assistance throughout the whole process.

Chananyu Pipatchatchawal

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

For several decades, various people have been interested and participated in many kinds of sports. Football, or soccer, is considered as one of, if not the most, famous sport on this planet. To be precise, there were four billion people watching football matches in 2020 [1]. This rising popular trend has increased the value of broadcasting rights, which contributes to a higher amount of prize pool for all competitions. Hence, from football clubs and fans point of view, winning as many matches as possible is desired and necessary for both economy and reputation purposes. At higher level, winning international competition, World Cup for example, also benefits the country [2]. Huge amount of prize pool and reputation growth awaits the champion. Hence, a higher number of winning matches will benefit both country and football club level.

Predicting football matches before it starts becomes one of the most challenging tasks among data scientists and researchers. This kind of project is also funded by some football clubs, as they might be able to prepare and select their players differently, maximizing the chance of winning for each individual opponent. Various existing research has tried to predict with different methods and algorithms, ranging from a complex neural network to basic logistic regression. It is possible that those results are improvable. Furthermore, numerical data gathered from FIFA, one of the most famous football video games in the past decade, is recently used as one of the input parameters to support prediction models.

English Premiere League (EPL) is the highest-level competition in England. It is one of the most popular football competitions in the planet. One of the reasons why people find the competition interesting is that it contains lots of unexpected results occurred every year. There is no guarantee whether top teams will beat lower teams. This happens for many years, and thus, makes this competitive league challenging for football match prediction. This thesis will aim to develop and experiments prediction model on this competition. This will help reflect out methodology on how it could perform on more varied and unpredictable matches.

1.1 Objectives

To create the hierarchical and ensemble classification model using several classification algorithms, which can predict result of football match.

1.2 Expected Outcomes

This thesis aims to propose a classification model for predicting the winner of football match. The model will be based on the line-up of players in each team. Thus, this methodology will be used as a guideline for the coach in selecting players for each match to increase the chance of winning.

1.3 Scope of the work

1. Match results of EPL's season 2015/2016 are predicted using 5 prior years as training data.
2. The proposed models are designed in two scenarios for both predicting results of three classes (Win/Draw/Lose) and two classes (Win/Lose), in order to be comparable with different existed papers.
3. Training features are based on overall rating of starting players from video games, attacking and defending rate, and recent matches result of each team.
4. All input features need to be pre-match attributes. Hence, in-game statistics, such as the numbers of fouls or shots, will not be used.

1.4 The Benefits of Research

This thesis proposed models can be used to assist football team in selecting starting players. This can one important guide to maximize the chance of winning for each team, based on each match opponent. Winning more matches will result in various benefit discussed earlier in the beginning of this chapter.

This main objective of this thesis is to develop a competing football match prediction model, that can forecast each match outcome based on eleven starting players of each team. It would be preferred if it can predict the future matches, not only finished matches. However, it is still such a difficult task as there are always surprises and no concrete formula in winning football matches.

CHAPTER II
LITERATURE REVIEW

As discussed earlier, multiple studies were conducted to solve this prediction problem, with the difference of some constraints and objectives. Some models focused on predicting matches beforehand, while some used all in-game statistics to predict the outcome. There are six related research papers included in this thesis.

Prasetio and Harlili used logistic regression to predict the exact English Premiere League (EPL) season, 2015/2016 [3]. The model was trained on five prior seasons, from 2010/2011 to 2014/2015, of the same competition. They tried using both four prior years and five prior years, with both including and excluding the testing year in training sets. Furthermore, only four parameters were used for training data, consisting of defending and attacking ratings for each side. Those numbers were collected from FIFA video games, one of the most famous football games licensed by the official football biggest organization, Federation Internationale de Football Association (FIFA). However, the authors tried to predict whether the winning team is the home team or away team only, leaving out the draw possibility. Their best proposed model was able to reach 69.5% accuracy in their experiments.

Similarly, logistic regression was also selected by Igiri and Nwachukwu [4]. They also applied Artificial Neural Network algorithm in building the prediction model. They tried to predict all three possibilities of football match outcome. However, the whole dataset used in their experiment was only one EPL season, 2014/2015. The accuracy achieved from their experiment was astonishingly high, at 95%. It is important to note that, besides using a small dataset, the authors also used in-game match statistics as input parameters. Those statistics were number of shots on target, corners received, yellow cards received, and many other numerical data occurred during the match. Thus, their proposed model focused more on classifying the winning team of finished matches based on match statistics, rather than predicting future match outcomes.

Snyder aimed to optimize both predicting football match result and betting strategy [5]. The interesting point to note is that, in this research, various non-football factors, stadium capacity, and traveling distance of the away team, for example, were added to assist player rating attributes. Logistic regression was selected for his model to predict all three match outcomes of EPL season 2011/2012, by training on one prior season. The model reached 51.06% accuracy with the two most important features being player evaluation and two previous matches.

Random forest classifier and multilayer perceptron model were investigated by Pugsee and Pattawong research [6]. There were two parts in their research. For the first part, they use two mentioned models to predict EPL season 2017/2018 result, using three prior seasons as the training dataset. Each classification algorithm will be used on three models, predicting 'home win' or 'home not win', 'draw' or 'not draw', and 'away win' or 'away not win'. They will select the better classification algorithm from this experiment to be used for the second experiment. All models with random forest classifier produced higher accuracy, with average at around 68.71%. Then, random forest classifier model then used to predict EPL season 2017/2018 season, with similar training and setting. The final results were at 79.09%, 81.81%, and 79.09% accuracy respectively.

Alfredo and Isa tried to compare the performance of multiple tree-based model algorithms on predicting football matches [7]. The authors studied C5.0, Random Forest, and Extreme Gradient Boosting algorithms. The experiment was set on using 14 independent features of ten EPL seasons, from 2007/2008 to 2016/2017. Dataset was also split using 10-fold cross validation for training and testing. Among those 14 input features, however, multiple in-game statistics were still included. The accuracies of all three proposed models were 64.87%, 68.55%, and 67.89% respectively.

Kumar constructed an in-depth analysis for football prediction [8]. In this research, they did not use any video games' numerical data for the predicting model, which is commonly used in other recent papers. Instead, they worked on raw match numerical data to construct other features. First, they built player rating models. They tried to predict player ratings from one to ten, using all match statistics, from just number of passes to even number of goals scored by each player. After that, they built football prediction model based on all in-game numerical data. Finally, they built future match prediction model, by combining the best models from the first two experiments. This final model used only pre-match features and tried to predict all three outcome possibilities. The author used four algorithms, including Bagging with Functional Trees, and AdaBoost with Functional Trees, Sequential Minimal Optimization (SMO), and Support Vector Machine (SVM). SMO produced the best result by using 27 input features and seven prior matches, with the accuracy of 53.3875%.

The summary of all related works studied can be found in Table 2.1. From all of the mentioned studies, each paper has some limitations in some areas. First, there is a huge gap difference in accuracy between using and not using in-game statistics. While in-game data does help classification tasks, it obstructs the concept of predicting future matches. Secondly, multiple studies did not account for all possibilities. Most of the time, authors tended to leave out ‘draw’ possibility, which is not preferred and appropriate if the goal is trying to predict actual football matches. This thesis desires to solve all above flaws, by constructing football match prediction model that used only pre-match attributes, and predict all plausible outcomes, including win, draw, and lose, with improved performance compared to existing methodology.

Table 2.1: Summary of all related works studied.

Paper	Using in-game Features	Classification Types	Best Methodology	Best Accuracy
Prasetio	✗	Win/Lose	Logistic Regression	69.51%
Igiri	✓	Win/Draw/Lose	Logistic Regression + Artificial Neural Network	95%
Snyder	✗	Win/Draw/Lose	Logistic Regression	51.06%
Pugsee	✗	Win/NotWin, Draw/NotDraw, Lose/NotLose	Random Forest Classifiers	81.81%
Alfredo	✓	Win/Draw/Lose	Multiple Tree-based Algorithms	68.55%
Kumar	✗	Win/Draw/Lose	Sequential Minimal Optimization	53.3875%

CHAPTER III

RESEARCH METHODOLOGY

This thesis research methodology can be divided further into four sections, including data collection, preprocessing, data partitioning, and classification models. This chapter will go through all sections.

3.1 Data Collection

There are two main parts of data used in this research, match data and video game's data. The whole data set was collected from the competition called European Soccer Database in Kaggle's website, a famous data science online community with a wide range of competitions [9]. The summary of the whole football match data is shown in Figure 3.1. In term of match data, six selected EPL seasons, from 2010/2011 to 2015/2016, were extracted from the huge dataset. There are 20 teams competing in each season, summing up to a total of 380 matches for each season. There are 115 match features, 42 player features, and 25 team features presented in the whole dataset. Those features includes both pre-match features, such as starting lineups, and in-game statistics. Furthermore, FIFA's video games ratings are used for both teams and players ratings. These data are processed to be appropriate for experiment usage in this thesis.

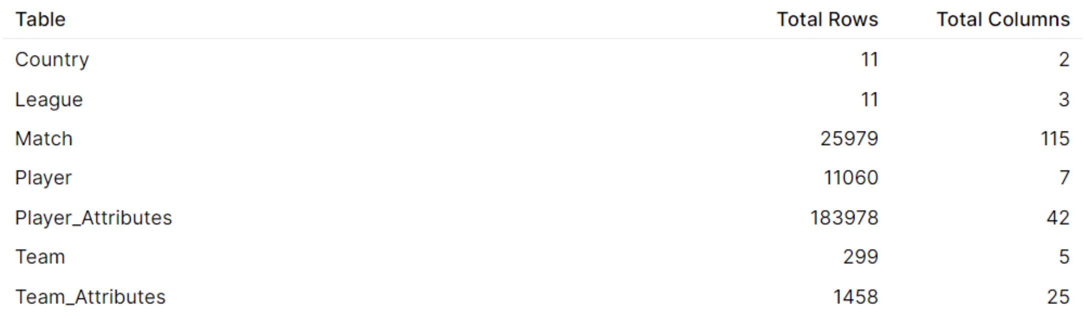


Table	Total Rows	Total Columns
Country	11	2
League	11	3
Match	25979	115
Player	11060	7
Player_Attributes	183978	42
Team	299	5
Team_Attributes	1458	25

Figure 3.1: Summary of whole football match data.

3.2 Preprocessing

From all the presented data, preprocessing will be done and categorize data into two main groups, including current match features and recent match history features.

Firstly, current match features are processed to help predicting football match results. This type of feature consists of match number, all teams' starting player ratings, and all teams' ratings. Match data is the match day of each season ranging from 1 to 38, in which the first one represents the first matchday, while the latter one being the last match day. Next, there are six rating features included to represent all starting twenty-two players, and five additional features for both goalkeepers. Lastly, nine features were included for both teams' styles of play. More details of current match features can be found in Table 3.1. Thus, this type of feature consists of 21 features in total.

For the second type of features called recent match features, 15 features are calculated to represent the recent form of each team. This type of feature can be divided deeper into three subgroups, including three latest results of home team against any other opponent, three latest results of away team against any other opponent, and the latest result of home team against away team. There is a total of five features representing each group, including the number of wins, the number of draws, the number of losses, the number of goals scored, and the number of goals conceded. All numbers will be average as each team, new team for example, might not have the same number of matches played. Table 3.2 shows the overall summary of recent match features created.

After getting all input features, each of them is normalized using min-max normalization, which forces all features to be in range of 0 and 1. This will help adjusting data values to common scale, while keeping the variance of the whole dataset.

Table 3.1: Current match features.

Feature Types	Each Team Quantities	Feature name	Data Type
Player	11	overall_rating	float
	11	potential	float
	11	sprint_speed	float
	11	reactions	float
	11	strength	float
	11	jumping	float
	1	gk_diving	float
	1	gk_handling	float
	1	gk_kicking	float
	1	gk_positioning	float
	1	gk_reflexes	float
Team	1	is_home_side	int(0/1)
	1	build_up_play_speed	float
	1	chance_creation_passing	float
	1	chance_creation_crossing	float
	1	chance_creation_shooting	float
	1	defence_pressure	float
	1	defence_aggression	float
	1	defence_team_width	float
	1	defence_defenderline_class	int(0/1)

Table 3.2: Recent match features.

Meeting type	Maximum number of matches	Parameter name
Home vs Any team	3	Average number of wins
		Average number of draws
		Average number of losses
		Average number of goals scored
		Average number of goals conceded
Away vs Any team	3	Average number of wins
		Average number of draws
		Average number of losses
		Average number of goals scored
		Average number of goals conceded
Home vs Away	1	Number of wins
		Number of draws
		Number of losses
		Number of goals scored
		Number of goals conceded

3.3 Data partitioning

For data partitioning, first, EPL season 2015/2016 is separated from all six interested seasons. This season will be called the final season for the rest of this thesis. This is the final season in the sense that it will act as the final test set for proposed models. This specific season is interesting because it contains many unpredictable matches. The winning champion of that season was at fourteenth place in the prior season. It was also the first time in the club's 132-year history to win this trophy. On the other end, the defending champion ended up in tenth place. Hence, predicting models would be promising if it could perform decently in this final season. On the other hand, five prior seasons, from 2010/2011 to 2014/2015, will be split using five-fold stratified cross-validation. Figure 3.2 shows data partitioning process on the whole dataset.

Five-fold means that all data will be split into five parts, with one part being a test set and the rest being a training set. Each part will behave as a testing set once. Thus, it will result in a total of five experimental sets, with 80% training and 20% testing data for each set. The concept of five-fold cross validation process can be found in Figure 3.3.

The Stratified algorithm is included to guarantee the consistency in the number of wins, draws, and losses matches proportion between all five data folds. An example of stratified dataset can be depicted in Figure 3.4. In this instance, three classes' ratios are 0.5, 0.2, and 0.3 for both the whole dataset and stratified dataset

All experiments in this thesis will use all different mentioned datasets, including five latest seasons, three latest seasons, and two latest seasons. This wide range of training sets will help in understanding how many seasons is optimal for each model to achieve highest performance.

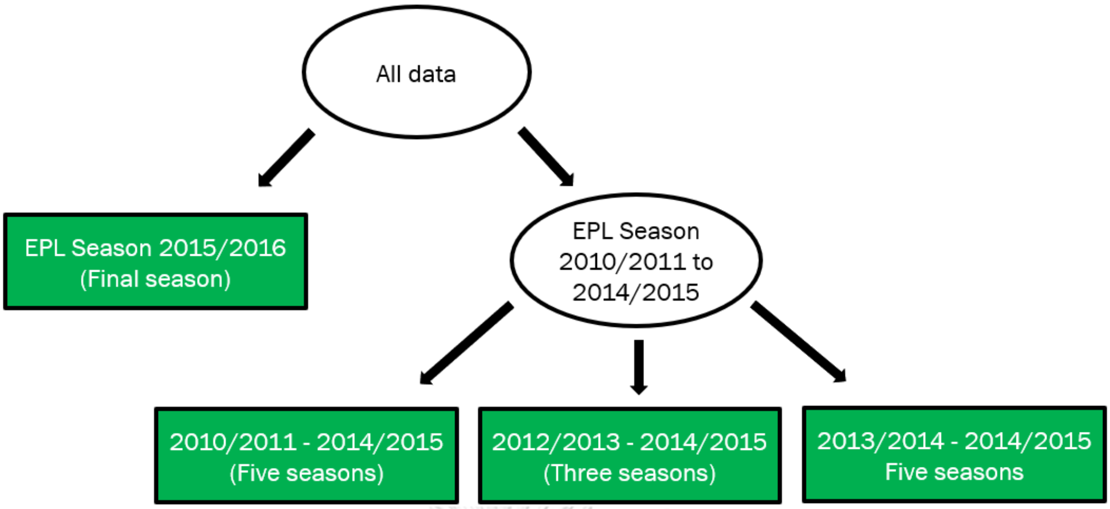

All data
EPL Season 2015/2016 (Final season)
EPL Season 2010/2011 to 2014/2015
2010/2011 - 2014/2015 (Five seasons)
2012/2013 - 2014/2015 (Three seasons)
2013/2014 - 2014/2015 Five seasons

Figure 3.2: Data partitioning process.

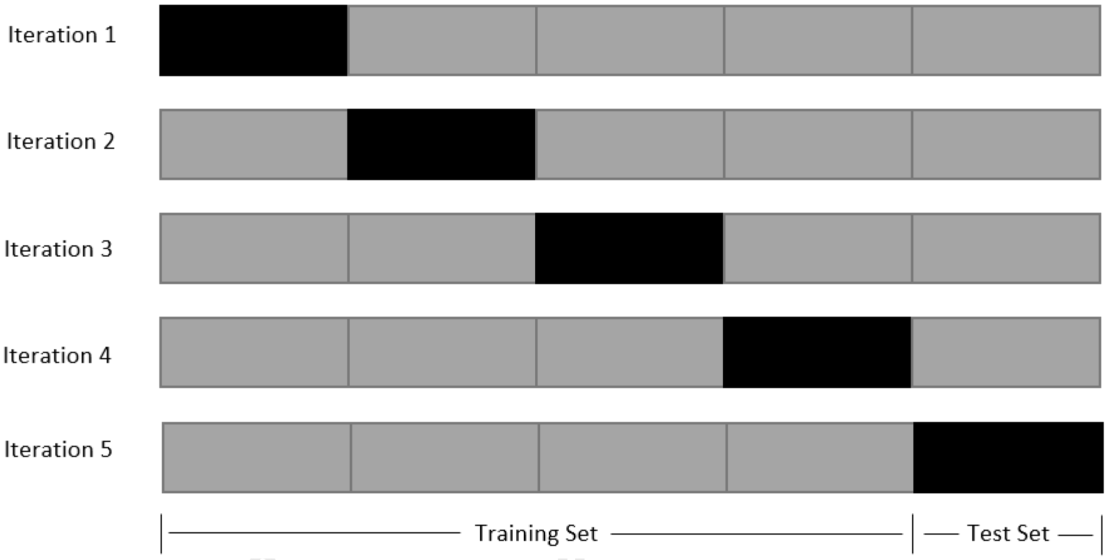


Figure 3.3: Five-fold cross validation data partitioning.

Grey block represents training data set, while black block represents testing set.

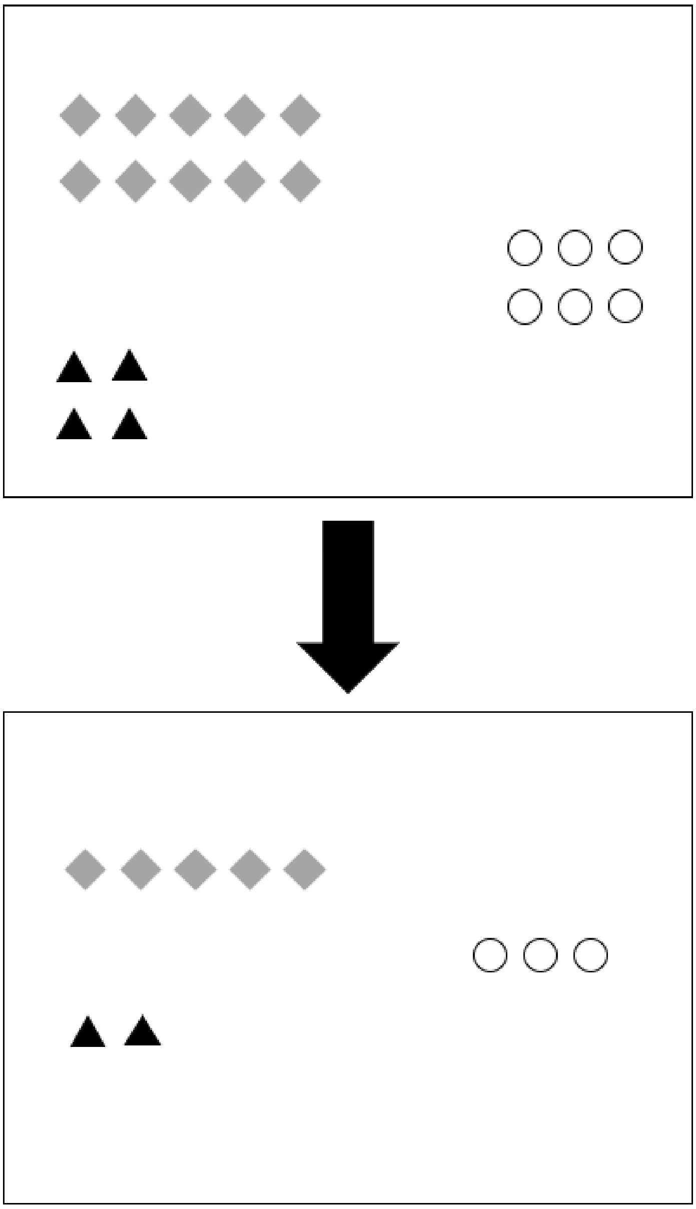


Figure 3.4: Example of stratified dataset.

3.4 Classification models

Six famous classification models are used in this thesis. All of them will be used to get the baseline accuracy, for comparison purposes. Six models selected are as follows

a. Multi-Layer Perceptron (MLP) is a feedforward learning type of artificial neural network algorithms. As presented in the name, more than two layers are composing the whole MLP model. Among all layers, at least one layer is a non-linear layer, which is called a hidden layer, which is placed between the first input layer and the last output layer. It is also applicable to non-linear activation functions, which make it capable of solving non-linearly separable data eventually. Figure 3.5 illustrates the sample architecture of MLP with two hidden layers.

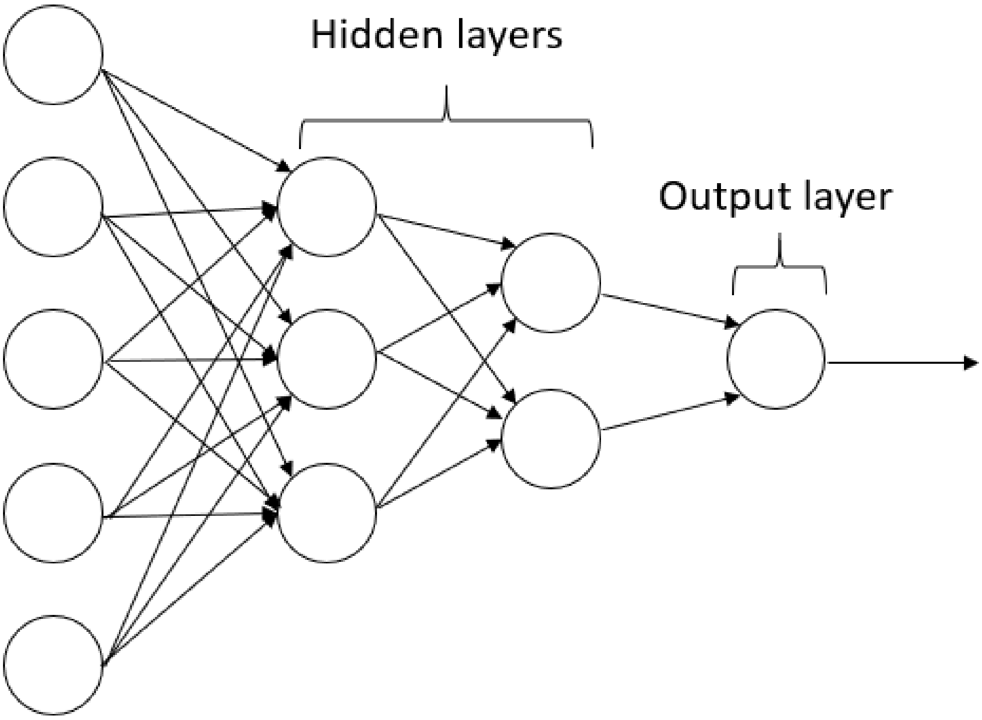


Figure 3.5: Example of MLP with two hidden layers.

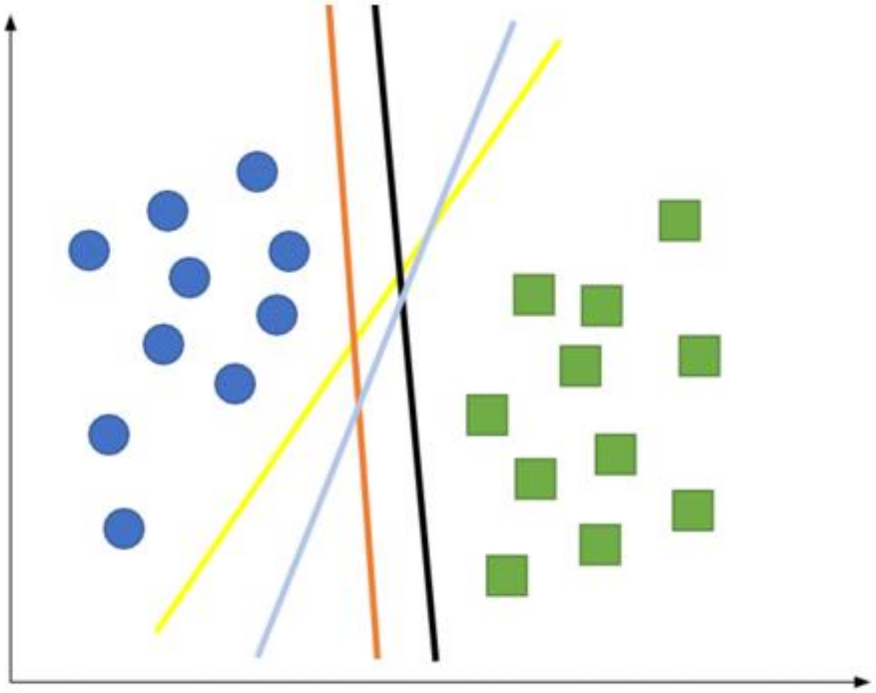


Figure 3.6: SVM concept with four hyperplane candidates.
In this example, yellow line would be the best hyperplane as it has the largest gap between each class and the plane itself.

b. Support Vector Machine (SVM) tries to classify data using the concept of hyperplane and dimensional space. Most of the time, input data will be mapped to higher dimensional space using a kernel function. The algorithm will then find the best decision hyperplane that can classify input data respected to their classes. This optimal hyperplane is considered by the margin between different data classes and the hyperplane itself, the larger the better. Example of SVM concept with four hyperplane candidates is shown in Figure 3.6.

c. Gaussian Naive Bayes (GNB) algorithm is a classification model constructed by using Bayes’ theorem, a renowned probability concept. The Gaussian wording refers to the input assumption that data points are collected from Gaussian, or normal distribution.

d. K-Nearest Neighbors (KNN) method has the concept of distancing and majority votes. The method will first find the K nearest data points around the new presented data point. Final classification results will be based on those K data points, by considering the majority classes among all K data points. Figure 3.7 shows deploying KNN classification concept on new unknown data point.

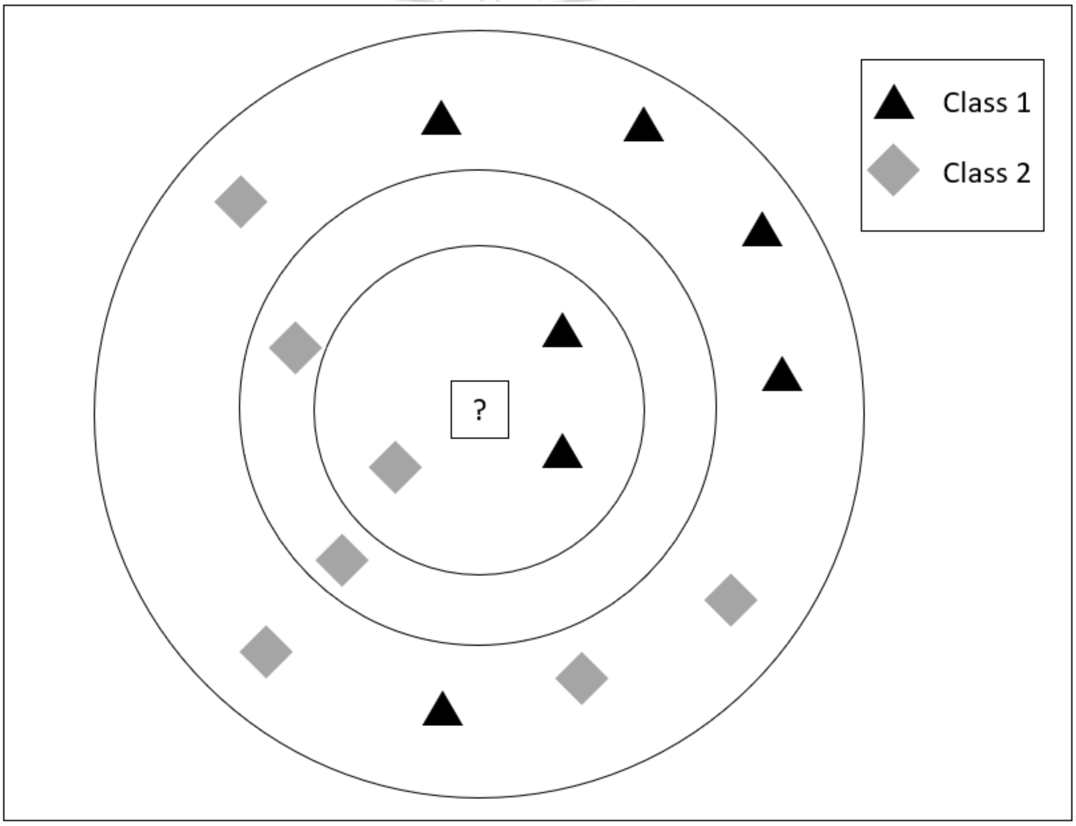


Figure 3.7: KNN concept.

The innermost circle represents running the KNN algorithm on new dataset with k equals to three, while the outermost layer represents the same process with k equals to 14.

e. Random Forest (RF) is a collection of multiple small decision trees. Each individual decision tree will try to make classification predictions itself. The final result can be decided by various criteria, such as the mode of prediction results, or average of prediction result from all existing subtrees, depending on what is desired for user’s usage. Figure 3.8 presents the sample architecture of RF classifiers.

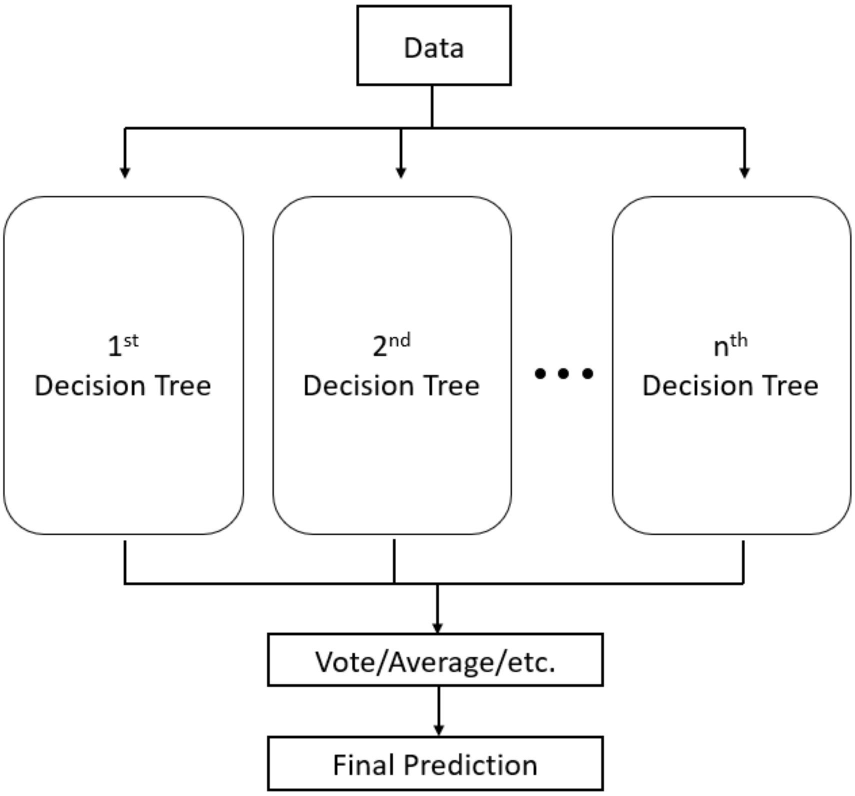


Figure 3.8: RF concept.

f. Gradient Boosting (GB) has some similarities to RF as both of them use the concept of combining multiple weak prediction models. The main difference between the two is that sub-models composing the main model in gradient boosting are combined at the beginning of the process's iteration. instead of combining at the end in RF. To be more precise, residuals or misclassification samples of each iteration are used to improve the main model's succeeding iteration. Sample GB concept is elaborated in Figure 3.9. In example case, if the new datapoint is x equals to six and y equals to four, it will be classified as triangle, circle, and triangle, after one, two, and three training iterations, respectively. Thus, it depends on how strict the training model is. If the model allowed two training iterations, it will be classified as circle, while being classified as triangle in the others. One point worth mentioning in this sample is that feature y is used twice, which is not possible in normal decision tree.

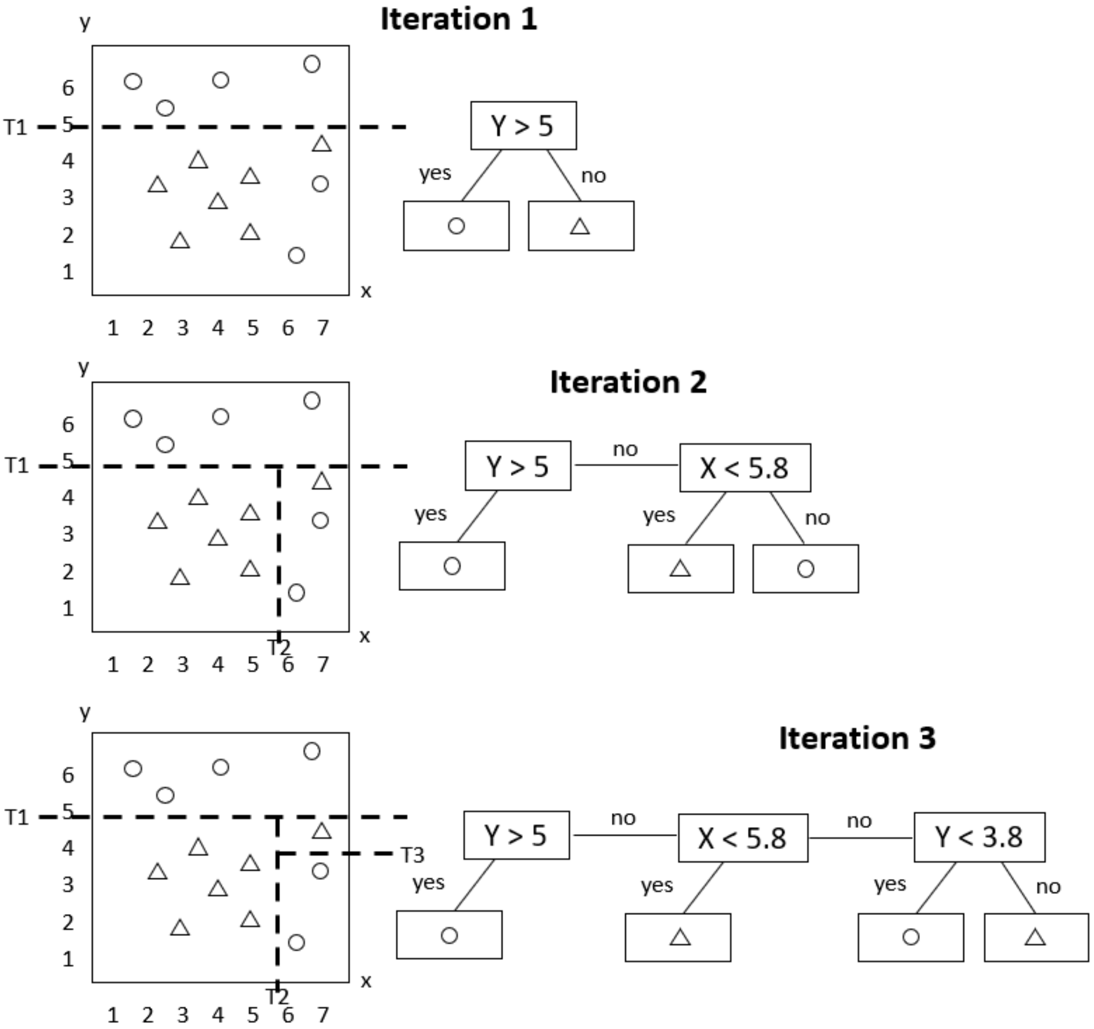


Figure 3.9: Sample of GB algorithm.

CHAPTER IV

PROPOSED METHODOLOGY

In this thesis, the proposed methodology brings the concept of fusion of classifiers to help improve football prediction models. The main insight of this concept is to combine multiple classifiers in specific ways, in order to solve more difficult tasks, which is predicting football match results in this scenario. Two fusion conceptual models are introduced in this thesis, including two Hierarchical models and one Ensemble model. Details of each model will be discussed in the following subsections.

4.1 Hierarchical Model

Hierarchical model connects multiple simple models in hierarchical fashion. Each individual model could be trained for different tasks, with similar or different datasets. It will ultimately connect to make final decisions at some stage in the hierarchy. Two unique hierarchical models are proposed, namely hierarchical model based on three classifiers, and hierarchical model based on two classifiers.

Firstly, hierarchical model based on three classifiers consists of three classifiers as the name suggests. For the same understanding, three classifiers will be called as classifier A, B, and C as demonstrated in Figure 4.1. Classifier A will predict whether each match is win or not win, while classifier B will focus on lose or not lose, and classifier C will try to predict win or lose only. Classifiers A and B will be trained on all available data. However, classifier C will extract only win and draw matches to be used as a training set, leaving out all draw matches. The concept behind this model is under the belief that, by training on data without draw results, that particular classifier, C in this case, should be better in predicting matches that tend to win or lose. Hence, if classifier A and B give win and lose prediction respectively, then that match will be subjected to win-lose classifier C. Otherwise, the result will be based on A and B, including win and not lose to be win, not win and lose to be lose, not win and not lose to be draw.

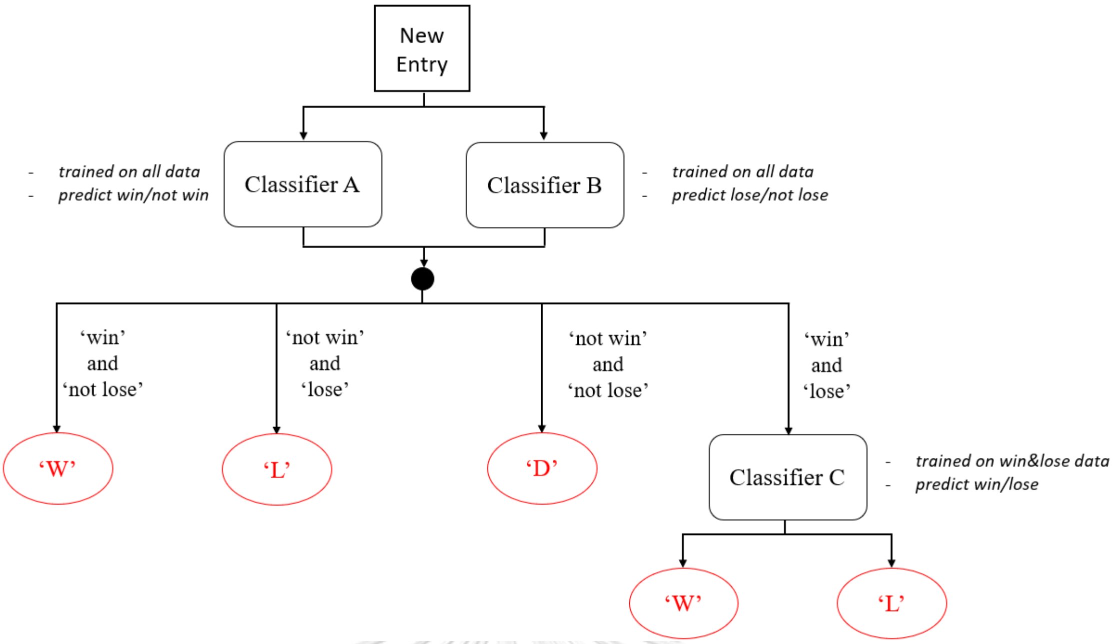


Figure 4.1: Architecture of hierarchical model based on three classifiers.

Secondly, hierarchical model based on two classifiers, A and B, is proposed and studied. In term of architecture, classifier A will focus on predicting whether the match is draw or not draw, with using all available processed data. On the other end, classifier B will be subjected without any draw matches, and thus, try to predict whether each match is win and lose. When the new data point is presented to this proposed model, it will first directed to classifier A. If classifier A predicts that match to be draw, then the process is concluded with draw as final result. Otherwise, that data point will be transmitted to classifier B to get the final prediction. This whole process is illustrated in Figure 4.2. This model is constructed under the hypothesis that, dividing 3-class classification task into smaller 2-class classification task, draw or not draw and win or lose, should make the whole task easier and improve the complete model accuracy.

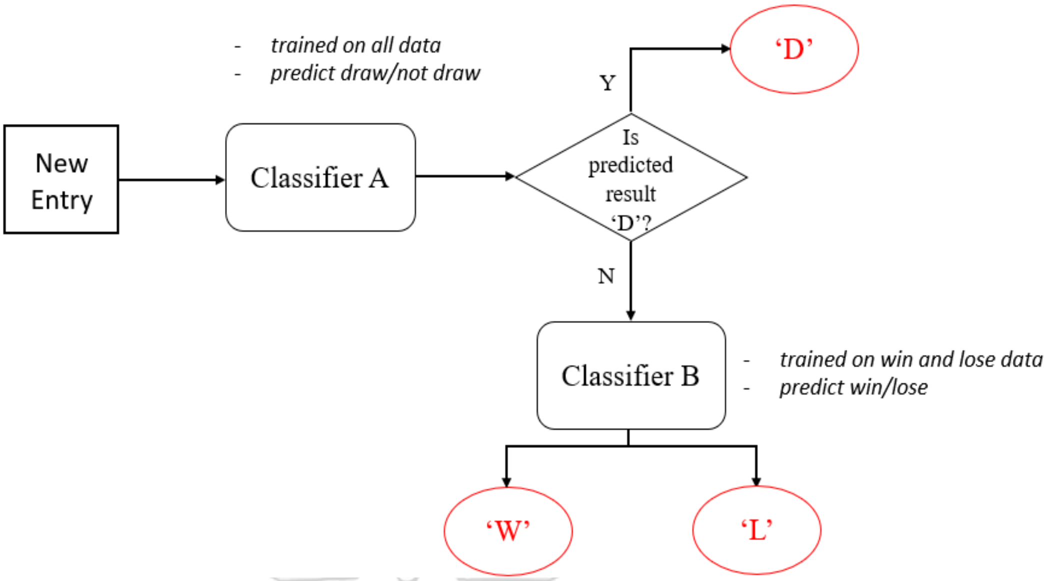


Figure 4.2: Architecture of hierarchical model based on two classifiers.

4.2 Ensemble Model

Ensemble model inherits the concept of a voting system to improve its accuracy. In this proposed model, multiple sub-models will be trained and make predictions individually. Final prediction will be made based on majority votes casted by all sub-models. Each sub-model will be different from each other and will have equal weights at the end.

There will be three different combinations of sub-model studied in this thesis. Each combination consists of three sub-models with different algorithms as shown in Figure 4.3. Combination composing the model will be selected by comparing the performance of each sub-model, which will be the result of all comparative models and hierarchical models. Selection results will be discussed in the next chapter.

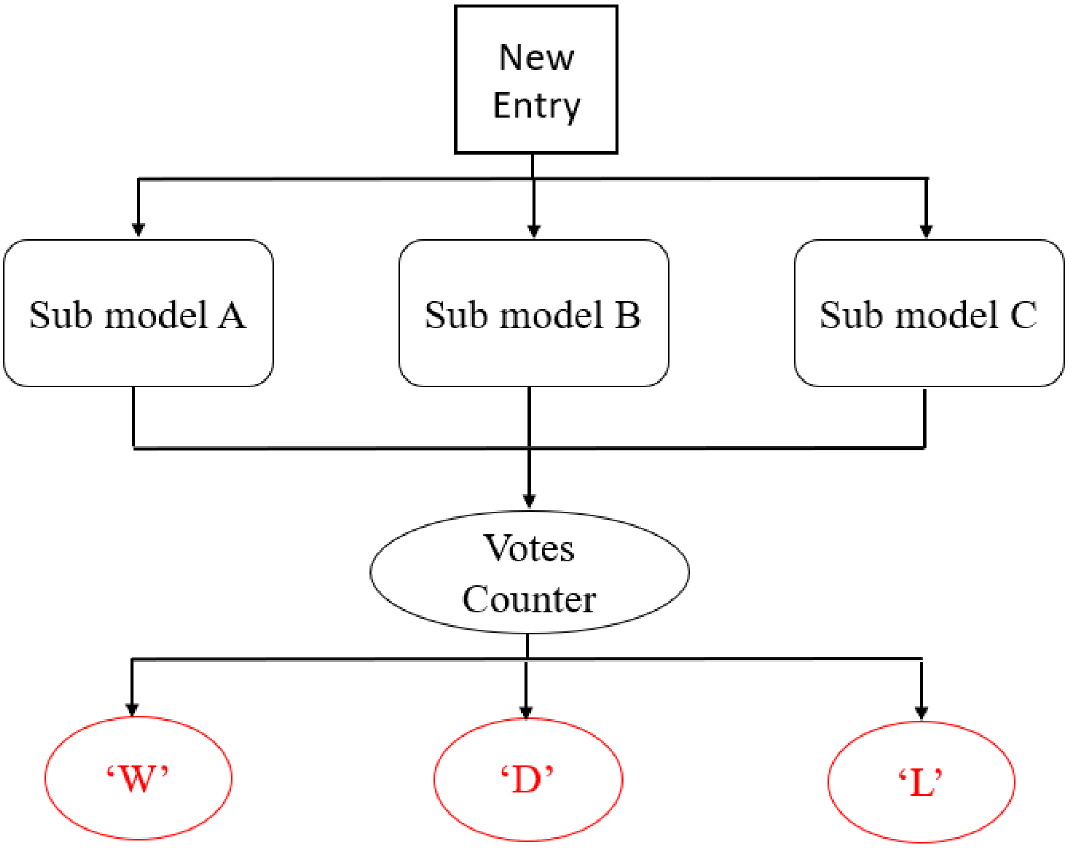


Figure 4.3: Architecture of ensemble model.

CHAPTER V
EXPERIMENT AND RESULT

Two experimental scenarios are constructed in this thesis. The first scenario is conducted to create the baseline performance for all algorithms. The result of this scenario will help setting up the second experiment, which aims to test all proposed fusion-based models. The details on all experiments are as follows.

5.1 Scenario 1

From the original dataset, the preprocessing process will produce three result datasets as discussed in chapter III, including five latest seasons, three latest seasons, and two latest seasons. Each dataset will be tested further with two feature sets. One set will use only recent match features, while the other use both recent match and current match features together. After pairing up each dataset with each feature type, all of them will be subjected to six comparative models mentioned in chapter III. In this experiment, both 2-class classification tasks and 3-class classification tasks will be tested on every classification algorithm.

In terms of the setting of each algorithm, each classification model will use the following configuration throughout this thesis. First, all MLPs will be created with using Adam solver, logistic sigmoid activation function, and maximum number of training iterations equals to 100. Secondly, SVM classifier with Radial Basis Function kernel will be trained. Third, KNN will be run with k equals to five. Next RF, all models will consist of 50 trees, and use entropy as a criterion to evaluate information gain. Lastly, GB models will use 100 training iterations. For GNB, however, does not required any initial parameter as the model can be constructed directly from training data.

Table 5.1 to Table 5.4 show the result of the first experiment. Table 5.1 presents the three-class classification accuracy of all comparative models using only recent match features, while the two-class classification accuracy of all comparative models using only recent match features is shown in Table 5.2. The result of using all features is illustrated in Table 5.3 for three-class classification task, and Table 5.4 for two-class classification task.

Table 5.1: Three-class classification accuracies from comparative models using recent match features.

Classifier	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
MLP	45.279	41.326	46.991	41.167	49.2	40.053
SVM	44.744	41.910	46.549	40.106	48.0	40.796
GNB	46.299	40.477	47.079	40.0	49.333	39.629
KNN	46.031	40.584	48.142	39.576	51.333	40.424
RF	39.647	40.424	42.832	41.379	44.267	39.522
GB	45.762	41.804	46.195	41.804	47.333	40.637

Table 5.2: Two-class classification accuracies from comparative models using recent match features.

Classifier	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
MLP	63.473	57.565	65.258	56.974	65.345	55.720
SVM	61.243	57.417	62.675	56.531	64.828	55.277
GNB	61.890	55.277	62.912	56.753	65.172	57.196
KNN	63.186	55.498	64.786	55.867	66.896	56.309
RF	58.792	55.351	61.386	58.303	62.241	53.432
GB	61.961	58.007	63.028	57.491	62.931	55.350

Table 5.3: Three-class classification accuracies from comparative models using recent match features and current match features.

Classifier	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
MLP	48.981	43.767	48.850	42.918	46.8	41.326
SVM	47.801	41.751	50.796	42.600	50.8	42.228
GNB	45.117	38.355	46.106	39.841	47.333	39.576
KNN	51.985	44.032	54.336	43.873	55.733	43.873
RF	49.303	43.077	51.593	42.599	52.8	42.759
GB	48.821	43.820	52.212	42.865	51.467	42.546

Table 5.4: Two-class classification accuracies from comparative models using recent match features and current match features.

Classifier	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
MLP	68.014	60.147	69.480	58.229	64.310	51.513
SVM	67.869	60.590	69.130	60.148	66.724	59.114
GNB	67.653	60.443	70.659	60.886	69.828	60.369
KNN	70.679	60.590	72.651	61.033	73.103	61.697
RF	69.526	61.550	70.071	60.664	68.276	60.590
GB	68.446	61.255	69.957	61.476	68.448	60.886

By using only recent match features, the highest accuracy achieved for the test set is at 51.333% by using two training seasons on KNN model for three-class classification task, and 66.896% using two training seasons on KNN model for two-class classification task. The lowest accuracy for the test set is by using the RF model with five training seasons at 39.647%, and 58.792% using the similar model and data, for three-class and two-class classifications, respectively. For the final season performance, the accuracy range for three-class classification is between 39.522% using RF with two training seasons, and 41.910% using SVM with five training seasons. As for two-class classification, the performance is between 55.277% from GNB model with five training seasons, and 58.007% from GB model with five training seasons.

On the other hand, using the combination of recent match and current match features results in improved performance on average. For test set, the accuracy range of three-class classification tasks increases to between 45.117% using GNB model with five training seasons, and 55.733% using KNN model with two training seasons. The accuracy gap of two-class classification objectives also increased to between 64.310% from MLP model with two training seasons, and 73.103% using KNN model with equal amount of training set. In the case of final season accuracy, the three-class classification accuracy reaches 38.355% by GNB model with five seasons of training set and 44.032% from KNN model with the same training data. Similarly, for two-class classifiers, the accuracy is at between 51.513% from MLP with two training seasons, and 70.679% using KNN with five seasons of training set.

From those results, it is obvious that using additional features tends to have a positive impact on prediction accuracy on average as expected. In terms of the optimal number of training seasons, using fewer seasons shows better accuracy for the testing set, while using all five seasons has more advantages for the final season. This statement is true to both three-class classification and two-class classification tasks, with higher gap difference in accuracy between test set and final season in three-class classification task.

Because of the mentioned similarities between three-class and two-class classification results, only three-class classification will be included for testing the further experiment. This decision can help minimize training and testing processes, while maintaining understandability of the whole process as two classification task produced similar trends. Furthermore, only a combination of two feature types will be considered in the next experiment. This is because using both of them is obviously better than using only one individually.

5.2 Scenario 2

In this scenario, all three proposed models, including two hierarchical models and one ensemble model, will be examined. This scenario will be split further into two experiments. The first experiment will try to test the performance of two proposed hierarchical models, while the second one will be designed for the proposed ensemble model.

In the first experiment, all processed data with all features will be subjected to both hierarchical models based on three classifiers and two classifiers. All mentioned classification algorithms will be tested. Within each trial, all classifiers composing the main hierarchical model will use the same classification algorithm. This will make each classification algorithm comparable with each other.

As presented in Table 5.5, the best classification algorithm composing the hierarchical model based on three classifiers is GNB. Accuracies are at 49.357%, 51.947%, and 52.267%, by using five latest seasons, three latest seasons, and two latest season training data respectively. Lower bound accuracies of the respected seasons, on the other end, are at 41.632%, 44.956 from KNN classifiers, and 45.067% from GB classifiers. In terms of final season accuracy, all models produce results lying between 37.719% using SVM with two training seasons, and 44.350% using GNB with three training seasons. From this average result, it is safe to claim that there is no significant improvement for the first hierarchical model, compared to previous comparative models.

Table 5.5: Accuracies from hierarchical models based on three classifiers.

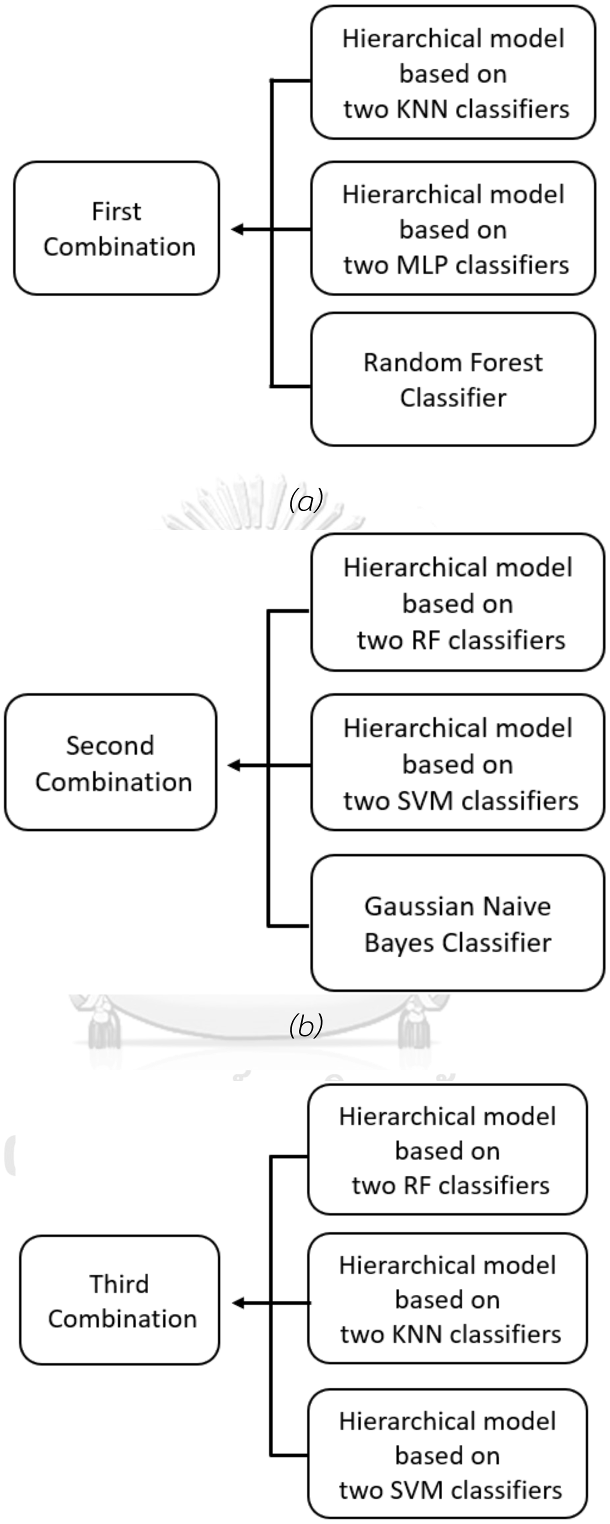
Classifier	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
MLP	44.261	38.196	47.522	39.523	47.600	39.416
SVM	43.830	38.515	46.195	38.780	46.0	37.719
GNB	49.357	43.130	51.947	44.350	52.267	42.600
KNN	41.632	38.143	44.956	38.833	46.933	38.409
RF	45.226	38.568	46.195	39.682	45.600	38.621
GB	43.619	40.212	45.929	39.257	45.067	39.682

In the second proposed hierarchical model, instead of using three classifiers, the model will try to predict football match results by using only two classifiers. The result can be depicted in Table 5.6. It has shown that the second hierarchical model performance has improved considerably. Besides GNB that performs relatively at the same level with the prior models, the accuracy ranges of the test set are between 49.196% using SVM and 52.200% using KNN for five training seasons, 50.442% using GB and 54.690% using KNN for three training seasons, and 51.467% using SVM and 56.533% using KNN for two training seasons. It can be seen that all but one classification algorithm has passed 50% accuracy in both three and two training seasons, which shows a very positive sign of the prediction performance. In case of the final season, other than the poor GNB performance, the accuracies are between 43.395% using SVM and 44.244% using MLP with five-season training data, 42.918% using GB and 44.032% using KNN with three-season training data, and 42.865% using SVM and 44.244% using KNN with two-season training data.

In order to select sub-model combinations for the final proposed ensemble model, previous experimental outcomes will be brought in for consideration by selecting some of the best performance models. Details of three chosen combinations are as follows. First combination includes hierarchical model based on two KNN classifiers, hierarchical model based on two MLP classifiers, and RF classifier as the first, second, and third sub-models. The architecture of the second combination is similar to the first combination, except using RF, SVM, and GNB as its respective classifiers. The third combination is different as all sub models are hierarchical models based on two classifiers. The first sub-model is a hierarchical model based on two RF classifiers, the second sub-model is a hierarchical model based on two KNN classifiers, and the third sub-model is a hierarchical model based on two SVM classifiers, respectively. The diagrams representing all three combinations of proposed ensemble model are presented in Figure 5.1.

Table 5.6: Accuracies from hierarchical models based on two classifiers.

Classifier	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
MLP	52.092	44.244	54.513	43.926	54.800	43.501
SVM	49.196	43.395	51.947	43.661	51.467	42.865
GNB	43.830	37.878	44.690	39.416	45.467	39.788
KNN	52.200	43.660	54.690	44.032	56.533	44.244
RF	49.197	44.138	53.186	43.077	52.267	43.342
GB	49.197	44.138	50.442	42.918	52.267	43.342


Hierarchical model based on two KNN classifiers
Hierarchical model based on two MLP classifiers
Random Forest Classifier
First Combination
Hierarchical model based on two RF classifiers
Hierarchical model based on two SVM classifiers
Gaussian Naive Bayes Classifier
Second Combination
Hierarchical model based on two RF classifiers
Hierarchical model based on two KNN classifiers
Hierarchical model based on two SVM classifiers
Third Combination

(a)

(b)

(c)

Figure 5.1: Three combinations of the proposed ensemble model.

a) The first combination. (b) The second combination. (c) The third combination.

Table 5.7: Accuracies of ensemble model.

Combination	Accuracy (%)					
	Five seasons		*Three seasons*		*Two seasons*	
	Test Set	*Final Season*	*Test Set*	*Final Season*	*Test Set*	*Final Season*
1	51.609	44.191	54.248	43.873	56.800	43.714
2	51.074	43.767	54.248	43.289	54.533	42.812
3	52.790	44.297	54.867	44.191	56.133	43.926

Accuracies of all combinations are presented in Table 5.7. In this proposed model, all combinations surpass 51% accuracy on the test set with all sizes of training set. The highest test set's accuracy achieved in this thesis is 56.800% by the first combination of the ensemble model with two-season training data. The highest accuracy reached for the final set, however, is only at 44.297%, which is not much distinct from the prior results. The average accuracy for the testing set is best using a two-season training set, and lowest with a five-season training set. It is reversed for final season accuracy, from the best using five training seasons, to two training seasons.

5.3 Football Experts Prediction

To get the better evaluation of proposed models, this subsection will compare our models against human intuition and experiences. Predictions on EPL season 2014/2015 and season 2015/2016 from one of the most renowned online pundit and expert for predicting accurate football match results, are collected for comparison purpose [10]. In the two mentioned seasons, the predictions were correct 190 and 178 out of 376 matches, respectively. This is equivalent to achieving the accuracy of 50.532% for season 2014/2015, and 47.340% for season 2015/2016.

From this number, this thesis proposed model is better in the regular season but being outperformed in the unexpected season. This might be because human started to side with the champion side. Furthermore, while player ratings in video games were updated weekly, those ratings for unrecognized players at the beginning of the season still could not reach the famed players’ level. Hence, football expert’s prediction could be more precise toward the end of this unexpected season.

CHAPTER VI
CONCLUSION

In summary, this thesis has proposed two types of fusion-based classification models, including hierarchical models, and ensemble model. For the first one, hierarchical models based on three classifiers and hierarchical models based on two classifiers are constructed. While the three-classifier version does not have great improvement, hierarchical models based on two classifiers has elevate accurate rates. This thesis has also investigated on the impact of using video games' ratings on football match prediction, which turns out to have significant benefits. In term of proposed ensemble model, three sub-model combinations are selected by considering prior experimental results. The peak performance is attained using the first combination of ensemble model. Hence, this thesis has produced two competitive football prediction models without using any in-game match statistics, which solve the future match prediction problem in multiple existing studies. The details are as follows.

First, using video games' ratings for all players and teams in current match features leads to better football match prediction performance. The accuracy range of all comparative models, between 39.647% and 51.333%, has increased to 45.117% and 55.733% for three-class classification tasks of the test set. This trend is also applicable to the final season performance but with lower gap, as the range has increased from between 39.523% and 41.910%, to 38.355% and 44.032%.

In the case of all proposed models, the first hierarchical model based on three classifiers did not have much improvement from comparative classification models. However, the second hierarchical model accuracy has significantly increased throughout all classifiers. The maximum accuracy reached from this model is at 56.533% for the test set, and 44.244% for the final season. Lastly, the ensemble model achieved the peak performance in this thesis at 56.800% using the first combination of sub-model with two-season training data for the test set, and

44.297% using the third combination with five-season training data for the final season.

In essence, the accuracies of two proposed models, including the hierarchical model based on two classifiers and the ensemble model, have enhanced remarkably. Furthermore, in terms of optimal number of training seasons, using all five latest seasons tends to be desirable for predicting unexpected outcomes in the final season, but not that great difference, while using fewer seasons is more beneficial for predicting the testing set. This might be due to the point that training five seasons might engulf more unexpected matches. Thus, trained models can perform better in the final season. On the other hand, fewer seasons is better for regular matches as it can better reflect on how each team has performed recently. The reason behind this remark might be because football teams rarely keep their playing performance steady through multiple years.

On the negative side, despite the improved accuracy in football match prediction models, it still does not result in significantly great performance, as the uppermost bound reached is only at 56.800%. The reason might be back to this thesis’s feature limitation as only pre-match features are accepted. Hence, on the bright side, the proposed model does surpass existing models with similar limitations, just not able to overcome models that use all possible attributes.

CHAPTER VII

FUTURE IMPROVEMENT

For future improvement, all proposed models could use feature selection and feature analysis with great benefits. The mentioned process could help in constructing proposed models in two ways. First, feature selection can definitely accelerate training processes. Although the current training process does not take extremely long time, a reduced number of features is still beneficial. Secondly, it might also help in elevating prediction accuracy. This might be possible if unnecessary features are detected and removed, resulting in sufficient training features which can also help prevent overfitting.

REFERENCES

[1] Shvili, J. The Most Popular Sports In The World. 2020; Available from: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.

[2] Staughton, J. What Benefits Does A Country Get After Winning The World Cup? 2020; Available from: https://www.scienceabc.com/sports/what-benefits-does-a-country-get-after-winning-the-world-cup.html.

[3] Prasetio, D. and D. Harlili, Predicting football match results with logistic regression*,* in Proceedings of the *2016* International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), p. 1-5. Penang, Malaysia. August, 2016

[4] Igiri, C.P., An Improved Prediction System for Football a Match Result. IOSR Journal of Engineering, 2014. 04: p. 12-020.

[5] Snyder, J. What Actually Wins Soccer Matches: Prediction of the *2011-2012* Premier League for Fun and Profit. Undergraduate Senior Thesis, Princeton University. 2013.

[6] Pugsee, P. and P. Pattawong. Football Match Result Prediction Using the Random Forest Classifier, in Proceedings of the *2*nd International Conference on Big Data Technologies. p. 154–158. Jinan, China. August, 2019.

[7] Alfredo, Y.F. and S. Isa, Football Match Prediction with Tree Based Model Classification. International Journal of Intelligent Systems and Applications, 2019. 11: p. 20-28.

[8] Kumar, G., Machine Learning for Soccer Analytics. Master of Science in Artificial Intelligence, option Engineering and Computer Science Thesis, KU Leuven. 2013.

[9] Mathien, H. European Soccer Database. 2016; Available from: https://www.kaggle.com/hugomathien/soccer.

[10] The BBC's Mark Lawrenson's Premier League football results predictions. 2016; Available from: https://www.myfootballfacts.com/stats/premier-league-by-season.

VITA

NAME	Chananyu Pipatchatchawal
DATE OF BIRTH	24 May 1996
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	B.Sc. Computer Science with Applied Mathematics minor, Mahidol University International College
HOME ADDRESS	46/23 Laddarom Village, Bangbon 4 road. Nongkhaem, Nongkhaem. Bangkok 10160