Faculty of FSIT
Department of Computing and Information System

Thesis On

“A Deep Learning Approach to Predict Football Match Result”

Course Title: **Project (Thesis)**

Course Code: **CIS499**

Submitted by:

Sumon Halder
ID: 201-16-508

Supervised by:

Mr. Md. Faruk Hosen
Lecturer
Department of Computing and Information System
Faculty of FSIT
Daffodil International University

Submission Date: 12 January, 2025

APPROVAL

This Project titled " A deep Learning Approach to Predict Football Match Result", Submitted by Sumon Halder, ID No: 201-16-508 to the Department of Computing and Information Systems, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing & Information Systems and approved as to its style and contents. The presentation has been held on 12-01-2025.

BOARD OF EXAMINERS

12.01.25

Md Sarwar Hossain Mollah — **Chairman**
Associate Professor and Head
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

12.01.25

Md. Nasimul Kader — **Internal Examiner**
Assistant Professor
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

12.01.25

Md. Mehedi Hassan — **Internal Examiner**
Lecturer (Senior Scale)
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

12.01.2025

Dr. Muhammad Shahin Uddin — **External Examiner**
Professor
Department of ICT
Mawlana Bhashani Science and Technology University

Declaration

I hereby declare that; this project has been done by me under supervision of **Mr. Md. Faruk Hosen, Lecturer**, department of Computing and Information System (CIS) of Daffodil International University. I am also declaring that this project or any part of there has never been submitted anywhere else for the award of any educational degree like, B.Sc., M.Sc., Diploma or other qualifications.

Supervised By

12-01-25

Mr. Md. Faruk Hosen
Lecturer
Department of CIS
Daffodil International University

Submitted By

12-01-25

Name: Sumon Halder
ID: 201-16-508
Department of CIS
Daffodil International University

ACKNOWLEDGEMENT

Firstly I express my thanks to almighty God for his divine blessing to make it possible to ready this paper successfully. I am also thankful to my supervisor Mr. Md. Faruk Hosen, Lecturer, Department of CIS, Faculty of FSIT, Daffodil International University. Profound Knowledge and keen interests of my supervisor in the field of deep learning impacted me to carry out this paper. His endless effort, guidance, continual encouragement, energetic supervision, valuable advice, many inferior drafts and correcting this at all stages have made it possible to complete this research paper. I would like to thank Mr. Md. Sarwar Hossain Mollah, Head, Department of CIS, Faculty of FSIT, Daffodil International University for his kind help to complete my Paper. Finally we would like to express our thanks to Pro. DR. Professor Dr. Syed Akhter Hossain Dean, CIS for giving us necessary information to complete this project report on time. After that we like to give thanks to our entire course mates in Daffodil International University, who took part in discussion while completing the course subject and course related work. Finally, we give cordial love and thanks to our beloved parents and friends for their mental support, strength throughout writing the project report in time.

DEDICATION

"To my dignified Parents and Teacher may they live long"

ABSTRACT

Football is the most watched and most played sport in the world. The 21st century there were approximately 250 million football players and over 1.3 billion people interested in football. Predicting the outcome of football matches has always been a topic of great interest among sports enthusiasts, analysts, and betting enthusiasts. With the rise of deep learning techniques and the availability of vast amounts of data, there has been an increased interest in developing accurate predictive models for football match results. This thesis presents a comprehensive study on using deep learning algorithms to predict the outcome of football matches. The goal is to leverage the power of deep learning models to improve prediction accuracy and provide valuable insights into the factors that influence match outcomes.

Table of Contents

List of Figures

List of Tables

Chapter: 1 Introduction

In this chapter, we are talking about the overview of sports analytics, background and significance of sports analytics, soccer data analysis and research problems and objects.
Now, we talk about these topic –

1.1 Background and Significance of Sports Analytics

Football is one of the most popular sports in the world. All people in the world love to watch and play football. Football is also known as soccer and played with a spherical ball. According to estimates, it is the most popular game in the world, with 250 million players across 150 countries. Football, like many other competitive sports, is now heavily reliant on data. So, that's why Research on football match outcome prediction is becoming more and more popular as a result of this trend. For football teams, being able to forecast the results of games is crucial since it gives them important information and a competitive advantage over rivals. It is both difficult and fascinating to predict the outcomes of sporting events. That means it's a difficult task to predict a football match result. But it's possible by using machine learning or deep learning to predict a match result. Machine learning and data collection methods, sports data analysis has grown more complex in recent years. One of the most popular sports in the world, soccer produces a lot of performance data that can provide important information about player efficacy, team tactics, and the factors affecting game results. In order to find important variables that could forecast game outcomes and uncover trends in team and individual performance, this study focuses on examining past English Premier League match data.
Sports analytics refers to the use of data analysis and advanced statistical tools to improve decision-making in the realm of sports. It has grown rapidly over the past two decades, moving from a niche practice to an essential part of modern sports management. Initially popularized by the book Moneyball, which detailed the data-driven strategies of Major League Baseball, sports analytics has expanded into various sports, including basketball, tennis, and soccer. Even though soccer is an unpredictable sport, historical data may provide important insights into trends and patterns that affect match results. According to earlier research, a team's chances of winning may be quantified by looking at metrics like possession, fouls, and shots on goal. For example,

although possession statistics by themselves may not necessarily indicate success, teams with a high percentage of shots on goal typically have a better chance of winning. Teams can develop more successful strategies and analysts can improve game outcome prediction models by knowing how these factors relate to one another.

The value of sports analytics resides in its capacity to give objective insights that supplement traditional scouting and coaching approaches. Analytics may be applied to player recruiting, injury prevention, team performance evaluation, and even fan interaction. Teams can obtain a competitive edge through data-driven tactics that are otherwise hard to see with the naked eye by measuring elements of training and games. The variety and complexity of data accessible have increased along with sports technology, making analytics a vital tool for anybody trying to maximize performance and enhance decision-making at all levels of competition.

1.2 Relevance of Soccer Data Analysis in the English Premier League

The English Premier League (EPL), one of the worlds most competitive and popular soccer leagues, is the subject of this study's analysis of historical match data. Because of the high level of competitiveness, diversity of play styles, and abundance of publicly available match data, the EPL provides a unique data environment. Data from previous EPL seasons will be analyzed in this study in order to pinpoint important variables that affect game results. Each game's final scores, team statistics (such as shots on goal, fouls committed, and corners earned), and officiating calls (yellow and red cards) are all included in the dataset. In order to develop predictive models that can anticipate the results of upcoming games, these parameters will be examined to see how they connect to match outcomes.

A more sophisticated comprehension of the game is made possible by the application of statistical and machine learning methods in sports analytics. To create prediction models, this study uses a variety of statistical analysis and machine learning methods. To determine which characteristics have the most effects on match results, techniques including regression analysis, decision trees, and ensemble approaches like random forests will be employed. The project intends to determine which variables—whether team-related (such as offensive and defensive stats) or external (such as referee tendencies)—hold the most prediction value by training models on historical data. After that, the accuracy of the model outputs will be assessed in order to produce trustworthy forecasts that can guide future match expectations.

Furthermore, the research's outcomes have wider ramifications for all parties involved in the soccer sector. Finding the primary factors that determine a game's outcome can help coaches and teams prepare for and modify their in-game tactics. For instance, teams may prioritize certain

foul patterns or a greater number of corners during practice if they know that these elements are associated with favorable results. In a similar vein, these insights may help sports analysts and broadcasters engage viewers with data-supported storytelling by offering deeper storylines and in-depth analysis during game broadcasts. Since data-driven predictions and insights offer another level of engagement and comprehension to their enjoyment of the sport, even spectators and fantasy league players may gain from these models.

We all know about the English Premier League (EPL) match fixer. If we watch EPL we know that every team played total 38 game. So, each team face a team twice and one home match and one away match. Sometime match result can be home team win, away team win or some time draw. The team with the most points wins the premier league, at the end of the season.

1.3 Research Problem and Objectives

The main research problem addressed in this paper is the identification of key factors that influence match outcomes in the English Premier League (EPL) and the development of predictive models based on historical data. While significant progress has been made in applying data analytics to soccer, many challenges remain, such as the difficulty of accounting for contextual factors (e.g., team form, match location, referee bias) and the inherent variability in match results.

The main objectives of this research are:

- To examine the Premier League's past match data in order to pinpoint important performance indicators that influence match results.
- To develop predictive models that can anticipate future match outcomes by utilizing a variety of machine learning methods.
- To make these models interpretable so that their conclusions may be usefully applied to actual situations, including performance reviews and changes to game strategies.
- To evaluate the prediction models' accuracy and limits and provide topics for development and more research in the soccer analytics field.

1.4 Structure and Overview of the paper

This essay is set up to walk the reader through a thorough investigation of the use of data analytics to soccer, with a particular emphasis on the English Premier League. The following is how the layout is made:

Introduction (Chapter 1): The introduction gives a broad overview of sports analytics, emphasizes the value of data analysis in the Premier League, and outlines the goals and challenges of the study.

Review of Literature (Chapter 2): summarizes the body of knowledge on soccer analytics, KPIs, and predictive modeling initiatives. It draws attention to information gaps that this study seeks to fill.

Methodology (Chapter 3): Explains the dataset, preparation procedures, and analytical techniques—such as machine learning and statistical analysis—that were utilized.

Experiment and result analysis (Chapter 4): Show the outcomes of the data analysis and model performance, including key insights gained from the analysis.

Discussion (Chapter 5): In this part, talking about how they affect soccer play and strategy, and contrasts them with previous studies. The study's shortcomings and some directions for further research are also covered in this section.

Conclusion and Future Work (Chapter 6): Provides a summary of the study's key conclusions and contributions and concludes by discussing the importance of data-driven methods in soccer analytics and also future plans of this paper.

Chapter: 2 Related Works

In this Chapter, discusses about various machine learning and deep learning to predict football match result.

2.1 Related Works

Using machine learning or deep learning techniques to make predictions in sports is often approached as a classification issue (1). Football is no exception, as most researchers aim to categorize each match into three outcomes: win, draw, or lose. Nevertheless, many scholars who have explored multiclass classification face the challenge of predicting draws accurately. The methodology outlined in an earlier work (2), where the author proposes using least squares to predict the outcomes of football and basketball games, represents one of the initial techniques introduced in the research literature for forecasting sports results. Since then, this topic has garnered several contributions. In (11), we see that their long short-term memory model failed to predict any draws, on the other hand (3) a conclusion also echoed by another study using a logistic regression model.

In (7), the author presents a model called pi-football, which utilizes a Bayesian Network to forecast the outcomes of matches in the English Premier League (EPL). This model incorporates both objective and subjective information while also considering the uncertainty of the available data. However, aside from a few notable exceptions, the findings from the early 2000s tend not to be very promising. Another article (11) discusses a convolutional neural network-based approach for predicting basketball game outcomes, highlighting that the convolutional layer enables the model to leverage player-level data. The authors also point out that aligning their predictions with those from bookmakers does not significantly enhance the accuracy of the final outcomes. Furthermore, (9) the study examined the potential of Artificial Neural Networks (ANNs) by analyzing various ANN-based methods for predicting sports results and identifying the challenges that still need to be addressed.

In [4], the author utilizes FIFA ratings to assess the football skills of various teams. These ratings are numerical figures developed by EA Sports that denote the capabilities and competencies of

players and teams. Building on Kahn's approach [13], this research achieved improved accuracy by analyzing a substantial dataset of 208 matches from the 2003 NFL season. This study highlights several factors, such as indicators for the home and away teams. The system presented in work [17] proposed a prediction model for forecasting the results of English Premier League (EPL) matches by incorporating both objective and subjective data, including team strength, current form, psychological factors, and player fatigue. Rana et al. [16] introduced a Logistic Regression model to anticipate match outcomes (home vs. away) in the English Premier League, utilizing SVM, XGBoost, and Logistic Regression classifiers for initial data classification and subsequently identifying the most effective algorithm for accurate labeling. The classifiers were applied to actual team data sourced from football data.co.uk, covering the seasons from 2003-04 to 2018-19, achieving a prediction accuracy of 65.63%. In [14], it was shown that while victories and defeats significantly reflect a team's performance during a match, predicting draws poses a challenge for machine learning techniques.

As opposed to, [15] compiled a dataset to forecast EPL match outcomes (home win, away win, or draw) by web scraping team ratings from FIFA, while considering each team's home and away performance. Their dataset encompassed FIFA ratings alongside performance data from the last ten seasons. They employed three machine learning classification techniques: Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF), achieving the highest accuracy of 59% with the SVM method. In [18], researchers trained their model on home and away wins, with an accuracy of 69.5% using data from five seasons. [20] Pointed out that only 2 of the 12 features used were non-engineered, indicating the importance of feature engineering in football match predictions. In [19], the authors calculated aggregated match statistic indicators for home and away teams based on past performances by taking either averages or sums. In [21], Alfredo et al. examined football match predictions using tree-based models, including C5.0, random forest, and extreme gradient boosting. They utilized a backward wrapper method for feature selection to enhance model accuracy. This study analyzed ten seasons of EPL match history with 15 initial features to predict match outcomes (home win, away win, or draw). The random forest algorithm achieved the best accuracy at 68.55%, while C5.0 recorded the lowest at 64.87%, and extreme gradient boosting provided an accuracy of 67.89%. In [23], logistic regression was applied to predict outcomes for the 2015/16 English Premier League season, focusing only on the possible victories or losses of the home team and not accounting for draws. This model reached a hit rate of approximately 69.5%, identifying that home and away team defensive records were the most significant influencers on predictions. In [22], the research explored the application of machine

learning to forecast football game outcomes based on match and player characteristics. A simulation study covering all matches in the top five European football leagues and their corresponding second divisions between 2006 and 2018 indicated that an ensemble strategy yielded statistically significant returns of 1.58% per match.

2.2 Summary

It is argued in the literature that soccer analytics are becoming more complex and essential in establishing match insights and performance prognosis. Some may include; enumeration of influential KPIs, improvement on the methods of learning such as machine learning, popularity of the expected goals models to measure goal-scoring opportunities.

There is no denying the low-scoring sport of soccer, the limitation of contextual information, and difficulties in interpreting a higher-level machine learning model. The main point of this study is: using advanced predictive techniques and focusing on model interpretability and usability. As such, this research narrows its investigation to the English Premier League, offering a study specific to one of the most popular soccer leagues globally and one of the most competitively contested leagues in the world which allows this research to fill gaps that exist in the current literature on soccer analytics.

Chapter: 3 Methodology

In this chapter, we discuss the proposed methodology and how this is accomplished. In this chapter we can see the main contribution of this project. Firstly, we figure out and discuss the architecture of the proposed model. Then, we discuss some preliminary steps and an overview behind that methodology. Later, we discussed the importance of the model elaborately with necessary pseudocodes.

3.1 Dataset Description and Source

In this research, we use 20 EPL (English Premier League) season, starting from 2005-2006 till 2024-2025. Also this data is historical match data from the English premier League (EPL). The football-data.co.uk [6] provides historical football matches statistics and soccer betting odds data. I have taken the dataset from this site.

Table 3.1: Distribution in the dataset (Target variables)

Home Team for match outcomes	Numbers of Records
Home Wins (H)	3,166
Away Wins (A)	2,068
Draws (D)	1,665
Total	6,899

The dataset also includes various information on team performances, matches outcomes, referee decision, players statistics, HS, AS, HF, AF, HTHG, HTAG, HR,AR etc.

3.2 Data Features and Variables Overview

In our dataset, there are many features to predict a football match result. And our dataset has three primary categories. Those are -

1. Team performance metrics
2. Match outcome variables
3. Referee/external factors

Those variables are connected by their relevance to soccer analytics to prediction football match results. Here we are discussing these three primary categories and how it's connected to our prediction.

3.2.1 Team Performance metrics

There are many team performance metrics in our dataset like Offensive Metrics, Possession and Passing Metrics, Set Piece and Defensive Metrics etc. Now we are talking about these Metrix.

In our dataset, we see many matrices like Home Team Shots, Away Team Shots, Home Team Shots on Target, Away team shots on target, Home Team Fouls Committed, Away Team Fouls Committed, Home Team Corners, Away Team Corners etc.

Home Team Shots (HS): Number of shots taken during the home matches.
Away Team Shots (AS): Number of shots during the away matches.
Home Team Shots on Target (HST): Number of shots on target taken during the home matches.
Away Team Shots on Target (AST): Number of shots on target taken during the away matches.
Home Team Fouls Committed (HF): During the matches, how many fouls committed by the home team.
Away Team Fouls Committed (HF): During the matches, how many fouls committed by the away team.

Home Team Corners (HC): How many Corners receive the home team during the matches.
Away Team Corners (AC): How many Corners receive the away teams during the matches.

3.2.2 Match Outcome Variables

Match Outcome variables represent the results of each match, which serve as the primary targets for prediction in this study. There are many match outcome variables in our dataset. Hare some match outcome Variables that we can be described:

Full Time Home Team Goals (FTHG): End of the match, how many goals scored by the home team.
Full Time Away Team Goals (FTAG): End of the match, how many goals scored by the away team.
Full Time Result (FTR): End of the match, how many goals scored by the home team and the away team and what is the result of the whole match. (Win, Draw or Lose).
Half Time Home Team Goals (HTHG): After 45 minutes (add extra time), how many goals scored by the Home team.
Half Time Away Team Goals (HTAG): After 45 minutes (add extra time), how many goals scored by the Away team.

So, it is true that these variables are very important to predict our match result.

3.2.3 External Factors

Some external conditions can significantly influence the dynamics and outcomes of soccer matches. There are many external factors that can change the match result. Here we discuss some primary External factors that are very important to predict our match result.

Referee Identity: During a Football match, Referees play a vital role throughout the whole match. Some wrong decisions can change the match result. So, Referee is an important external factor. If we see our dataset, several features that are connected to Referee Identity. These features are HR, AR, HY, AY, HF, AF etc. These feature results are created by a Referee decision.

Match Venue: Match Location is an important factor to predict a match result. If we know about the English Premier League (EPL), we see when a match is played against two teams and the

home team is benefited by many other factors. Captures the impact of home-field advantage, which has been shown to affect team performance.

Crowd Attendance: Home teams are benefitted by their Crowd attendance. There is no denying the fact that Crowd Attendance is an important factor to predict a football match result.

According to the above writing we can say that these factors ensure the model accounts for contextual influences beyond team performance alone.

3.3 Data Preprocessing Steps

In our dataset, we have 23 columns, covering various aspects of English Premier League (EPL) matches. Those aspects are match outcomes, Player performance, Referee Decision, Team name, Team performance ability etc. In this area, we are talking about Missing data and Feature Engineering. Firstly, we talk about missing data.

3.3.1 Identify Missing Data

Our dataset consists of 23 columns and a number of data records 6,899.Also, in this research comprises the EPL matches spanning 20 seasons, starting from 2005-2006 till 2024-2025. If we know about the English Premier League (EPL) we know that there are 38 matches played each season and there are 20 teams in the league each season and we have data for 20 seasons from 2005-2006 till 2024-2025 season. If we calculate our total dataset, we should have 7,620 data but we actually have 6,899 data. So, it's clear that some data is missing. The reason why this happened is that the EPL season 2024-2025 has some missing data. Because this season is going on now. So we were only able to take some data from this season.

3.3.2 Feature Engineering

In this area, we are talking about Feature Engineering. There should be no controversy over the fact that Feature engineering was undertaken to extract additional insights from existing columns and enhance the predictive power of the dataset. According to our dataset, some of the features are given below:

Match Outcome Variable Engineering:

- Match result encoding: Full Time Result (FTR) is the main target column in this study. So firstly, change the FTR column (Home Win, Away Win, Draw) into numerical categories. The numerical categories are: Home win = 1, Away win = -1, Draw = 0.

- Goal Difference (GD): Derived as FTHG - FTAG to measure team dominance in each match. Here, FTHG = Full Time Home Team Goals, FTAG = Full Time Away Team Goals.

Team Performance Metrics:

- Total Shots: Sum of HS (Home Shots) and AS (Away Shots) to think about the team attacking efforts.

- Player Shot Accuracy: Shot Accuracy is also an important feature to predict a match result. So, calculate as (HST/HS) for home and (AST/AS) for away, representing the percentage (%) of shots on target.

Home Location advantage Feature:

- Home Advantage Score: Count as (HS - AS) + (HC - AC) to evaluate the overall dominance of the home team in key areas.

- Away Disadvantage: Usually away teams always get some disadvantage. If we compare the metrics difference like AF, AY & AR with the home team then we analyze the difference.

External Factor Feature: External Factor Feature also is very important to predict a football match result. External Factor like match day, month and year that extract from the Date column. And then, create a Weekend Match binary feature to distinguish between weekend and weekday games.

Team Disciplinary Metrics:

- Count yellow (HY, AY) and red cards (HR, AR) for creating a new Total Card Feature to analyze how the team is balanced.
- Calculate the ratio of fouls committed to cards received (HF/HY, AF/AY) to assess refereeing trends.

Referee Impact:

- Encode Referee as a categorical variable for our model training.

3.4 Model Building and Testing

Here, we are talking about our six Machine Learning Model and after training our model, testing our dataset to see how they will do.

3.4.1 Model Building

Predictive models are constructed using different machine learning algorithms. Below written are a few of the popular algorithms for predicting a football match result.

1. Random Forest
2. Naive Bayes
3. Attention-Based CNN
4. Linear Regression
5. Support Vector Machine (SVM)
6. K-Nearest Neighbors (KNN)

These models are trained using the training dataset.

3.4.2 Model Testing

Test the model: After training the models are tested on some external testing dataset to see how well they do. Predictions are made, Performance metrics are calculated.

3.5 Performance Evaluation

To be sure about the accuracy and reliability, several metrics are used to evaluate the performance of the models. Here are these metrics-

3.5.1 Accuracy (ACC)

Measures the proportion of correctly predicted instances over the total instances.

3.5.2 Area under the Curve (AUC)

Evaluates the model's ability to distinguish between classes, providing a measure of discrimination.

3.5.3 Matthews Correlation Coefficient (MCC)

A robust metric for imbalanced datasets, assessing prediction quality in a balanced manner.

3.5.4 Precision and Recall

Precision measures the accuracy of positive predictions, while recall assesses the ability to capture actual positives.

3.5.5 Sensitivity and Specificity

Sensitivity refers to the true positive rate, and specificity indicates the true negative rate, offering a balanced evaluation.

3.5.6 F1-Score

Combines precision and recall into a single metric, providing a harmonic mean that balances false positives and negatives.

3.6 Tools and technologies

In this study, we use different types of tools and technologies. The study we implemented in python. We write code in the Jupyter Notebook. Here are some libraries that is used in this study-

- TensorFlow
- Scikit-learn
- Pandas and Numpy
- Matplotlib

3.7 Summary

In this chapter, firstly we talk about our data source and data description. Then, we talk about data features and variables. Talking about our six machine learning modes and training these models to realize how well our data is. If we see this chapter, we also know about feature engineering, missing data, model building and model testing, performance evaluation, tools and technology and some other things.

Chapter: 4 Experimental Setup

In this chapter we are discussing our experimental setup. Discussing dataset splitting, sampling techniques, classification models and model evaluation in this chapter. To predict football match results, we use five machine learning models and one deep learning model. And our target column FTR (Full Time Result) is ready to predict match results, using features from the dataset.

4.1 Sampling Techniques

We use 20 ELP (English Premier League) season for this research. Also this data is historical match data from the English premier League (EPL). The dataset is split into training and testing subsets. For training we used 80% data from our dataset and for testing we used 20% data for our research.

After training our model, our model is ready to predict the football match result. We know that our target column is FTR (Full Time Result). So, it's a multiclass prediction because our model predicts “Win”, “Draw”, “Lose”.

4.2 Classification Models

Here we discuss our six machine learning model. Every model predicts match results in their own way. Now we discuss about these models-

4.2.1 Random Forest

Nobody can make an argument over the fact that the Random Forest (RF) classifier is a type of ensemble learning algorithm that merges the predictions from various decision trees to arrive at a conclusion. We know that each decision tree is constructed differently and trains on a subset of the data, which is obtained using a method known as bootstrap sampling. There is no denying

the fact that when the RF classifier makes predictions, each tree independently estimates the class label for a given data instance, and the ultimate prediction is the class that receives the most votes.

4.2.2 Support Vector Machine (SVM)

Firstly we can say that Support Vector Machine (SVM) is also utilized for predicting football match outcomes. Support Vector Machine is a supervised machine learning technique applied to both classification and regression tasks. We all know that SVM identifies the best hyperplane to distinguish between different classes by maximizing the distance between the nearest points (support vectors) of each class and the hyperplane. Normally, SVM minimizes classification errors through a soft margin strategy, which strikes a balance between accuracy and generalization.

4.2.3 K - Nearest Neighbors (KNN)

We all know that, K-Nearest Neighbors (KNN) is a general and non-parametric machine learning algorithm that is used for both classification and regression tasks. In the field of football, KNN predicts match outcomes by investigate the K nearest data points in the feature space based on distance metrics like Euclidean or Manhattan. Nobody can make an argument over the fact that KNN uses majority voting for both classification or averages for regression. KNN requires feature scaling and is mainly used for smaller datasets.

4.2.4 Attention Based CNN

Attention-Based CNN is a deep learning model and it consolidates convolutional neural networks (CNNs) with attention mechanisms to focus on significant features in the data. We all know that attention-based CNN models improve standard convolutional neural networks by incorporating attention mechanisms that help emphasize the most significant features within the input data. These mechanisms, which include spatial, channel, self-attention, and hybrid types, enable models to dynamically prioritize key areas and attributes. Frameworks such as SE-Net, CBAM, and DANet implement these methods to achieve leading-edge performance in areas like image classification, object detection, semantic segmentation, and biomedical imaging. Although these models encounter issues like increased computational demands, innovations in efficient and interpretable attention mechanisms are expected to expand their potential uses even further.

4.2.5 Linear Regression

We all know that, Linear Regression is an algorithm used in supervised learning to forecast continuous results by analyzing the connection between a dependent variable and one or more independent variables. By minimizing the Cruel Squared Blunder (MSE), it decides the best-fit line to foresee persistent results. There is no denying the fact that this model's ease of use and clarity in interpretation render it useful for tasks such as analyzing trends and making predictions. Linear Regression performs well with smaller datasets; however, it has limitations due to its underlying assumptions of linearity, its susceptibility to outliers, and difficulties associated with multicollinearity when dealing with multiple regression situations.

4.2.6 Naive Bayes Classification

Nobody can make an argument over the fact that Naive Bayes is a classification algorithm that relies on probabilities, grounded in Bayes' Theorem, to determine the likelihood of a particular class based on available data. It operates under the premise that the features are independent, which simplifies the calculations while ensuring efficiency. We all know that Variants like Gaussian, Multinomial, and Bernoulli Naive Bayes handle continuous, count-based, and binary data, respectively. Known for its simplicity and motion. Naive Bayes is widely used in text classification, spam filtering, sentiment analysis, and medical diagnosis. Although it works well for many applications, it may face challenges with features that are highly related and issues of zero probability. These problems can be alleviated through the use of techniques.

4.3 Model Evaluation

In this task, the evaluation of all models used in our experiment is conducted through various metrics like Accuracy, F1 score, Area under the ROC curve (AUC), Matthews Correlation

Coefficient (MCC), Recall, Precision, Sensitivity, and Specificity. We know that, accuracy is defined as the ratio of correct classifications to the total number of instances in the dataset.

$$accuracy = \frac{TP + TN + FP + FN}{TP + TN}$$

Additionally, the F1 score serves as a metric that maintains the precision and recall of a model. There is no denying the fact that precision assesses the proportion of true positive predictions among all positive predictions made by the model, while recall measures the proportion of true positive predictions against all actual positive instances in the dataset.

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

F1 Score = 2.$\frac{Precision\,.Recall}{Precision + Recall}$

It is widely accepted that the AUC is often used to evaluate the performance of various classification models. This metric captures the trade-off between the true positive rate and the false positive rate across different classification thresholds by plotting the ROC curve. On the other hand the AUC score represents the area under this curve, with a higher AUC indicating better performance in distinguishing between positive and negative classes. In multiclass classification scenarios, the “one-vs-rest” approach is employed.

Sensitivity and recall are synonymous terms; sensitivity refers to the model's capacity to accurately identify positive instances.

Sensitivity = $\frac{TP}{TP+FN}$

Specificity gauges the percentage of actual negatives that the model correctly identifies, reflecting its capability to minimize false positives.

Specificity = $\frac{TN}{TN+FP}$

Matthew correlation coefficient (MCC) is a metering of the quality of a binary classification. MCC works for four matrices (TP, TN, FP, FN). And return value this way: 1 to 1, where 1 shows a well prophecy, 1 shows a perfect mismatch between reality and prediction, 0 represents a random prediction.

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)\ (TN+FP)(TN+FN)}}$$

Chapter: 5 Result Analysis and Discussions

In this chapter, we are talking about our model results. In this research, we use six machine learning model and gave excellent performance. So, the approach taken to test different combinations of variables yielded good results. Various model results and metric result discussed below -

Firstly we are talking about our model result. In this paper, we are talking about our six machine learning model. Analyze all the model results and choose the best model for match prediction. And our target model is attention Based CNN model. The results of different models are shown in the table below -

Model	ACC	AUC	MCC	Precision	Recall	Sensitivity	Specificity	f1
Randomforest	0.992%	0.999%	0.987%	0.992%	0.992%	0.992%	0.997%	0.992%
Naive Bayes	0.750%	0.898%	0.616%	0.770%	0.750%	0.731%	0.878%	0.757%
Attention based CNN	0.951%	0.978 %	0.925%	0.942%	0.953%	0.953%	0.969%	0.947%
Linear regression	0.897%	0.999%	0.750%	0.874%	1.0%	1.0%	0.643%	0.933%
SVM	0.997%	0.99%	0.995%	0.996%	0.996%	0.996%	0.997%	0.996%
KNN	0.731%	0.860%	0.58%	0.729%	0.731%	0.731%	0.86%	0.727%

Table 5.1: All Models results

In the above table, there are six machine learning models and their results. If we analyze the results then we see Random Forest, Linear Regression, and SVM stand out due to their consistent high performance across all metrics. KNN and Naive Bayes did not perform well that's why these two models are not suitable for this task. But the Attention Based CNN model gives a strong performance. So, to predict a football match result, we use the Attention Based CNN model.
Now we discussed about result broadly and various metrics -

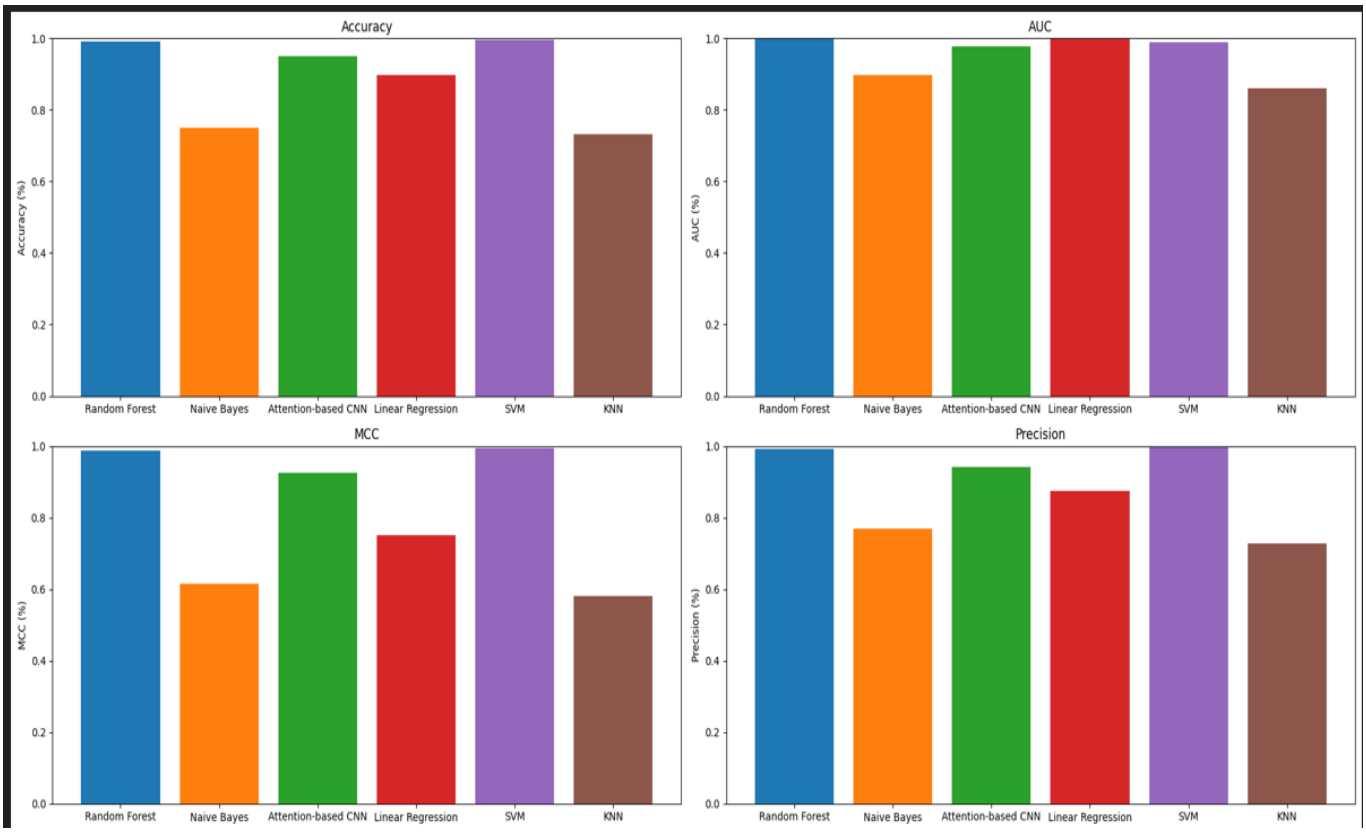


Figure 5.1: Bar chart of Accuracy, AUC, MCC, precision.

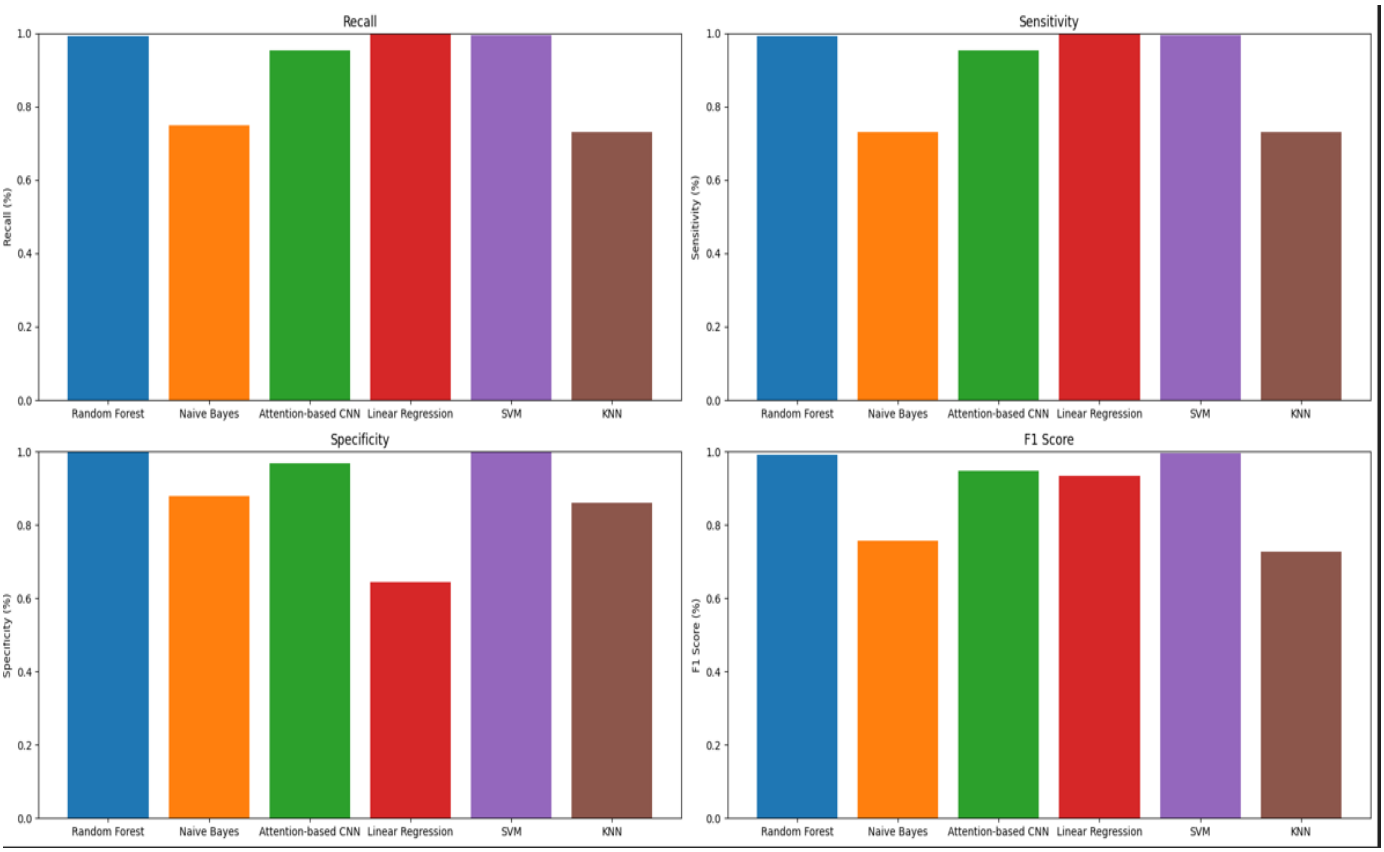


Figure 5.2: Bar chart of Recall, Sensitivity, Specificity, F1 score.

If we see our Bar chart, Figure 5.1 shows model Accuracy, AUC, MCC, Precision and Figure 5.2 shows Recall, Sensitivity, Specificity, F1 score. Now we discuss these metrics. Firstly, we discussed accuracy metrics and we see Random Forest, SVM, Linear Regression, and Attention-based CNN perform comparably well, while Naive Bayes and KNN have lower accuracy.

AUC is almost similar to Accuracy. Here, Random Forest, SVM, Linear Regression, and Attention-based CNN perform comparably well, while Naive Bayes and KNN have lower accuracy.

If we see Figure 5.1 and Figure 5.2 then we see high MCC values are seen for Random Forest, SVM, and Attention-based CNN, whereas KNN shows weaker performance.

And reflects the proportion of true positives out of all positive predictions. Random Forest and SVM lead in this metric.

Other metrics Sensitivity, Specificity, Recall and F1 Score are also doing well.

According to the above analysis we can say that Random Forest and SMV give high performance and Naive Bayes gives low performance. But the outcomes of the Attention Based CNN model is perfect to predict a football match result.

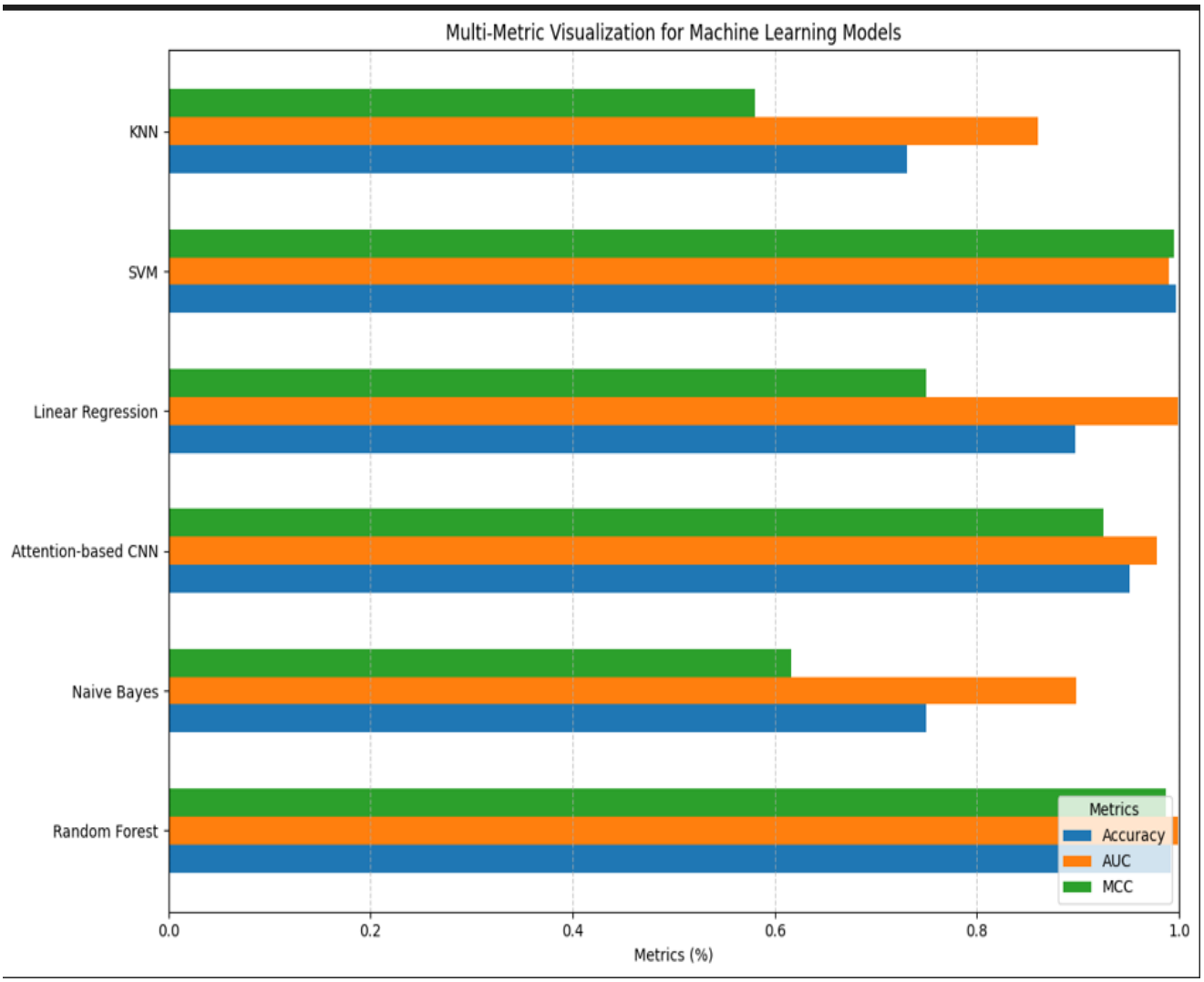


Figure 5.3: Multi-Metric Visualization for Machine Learning Models

Figure 5.3 represents a Multi-Metric Visualization for Machine Learning Models. Here, three main metrics: Accuracy, AUC, MCC and our six machine learning models. Metrics (%) represent model

performance. Higher values indicate better performance. Now we discussed our model performance and model metrics.

According to Figure 5.3:
Random Forest: High performance across all the metrics.
Linear Regression: Almost same Random Forest.
SMV: Strong and High Performance.
Attention Based CNN: Consistently good performance and all metrics result also well. This model is our target model to predict match outcomes.
Naive Bayes: Moderate performance.
KNN: Low performance across all the metrics.

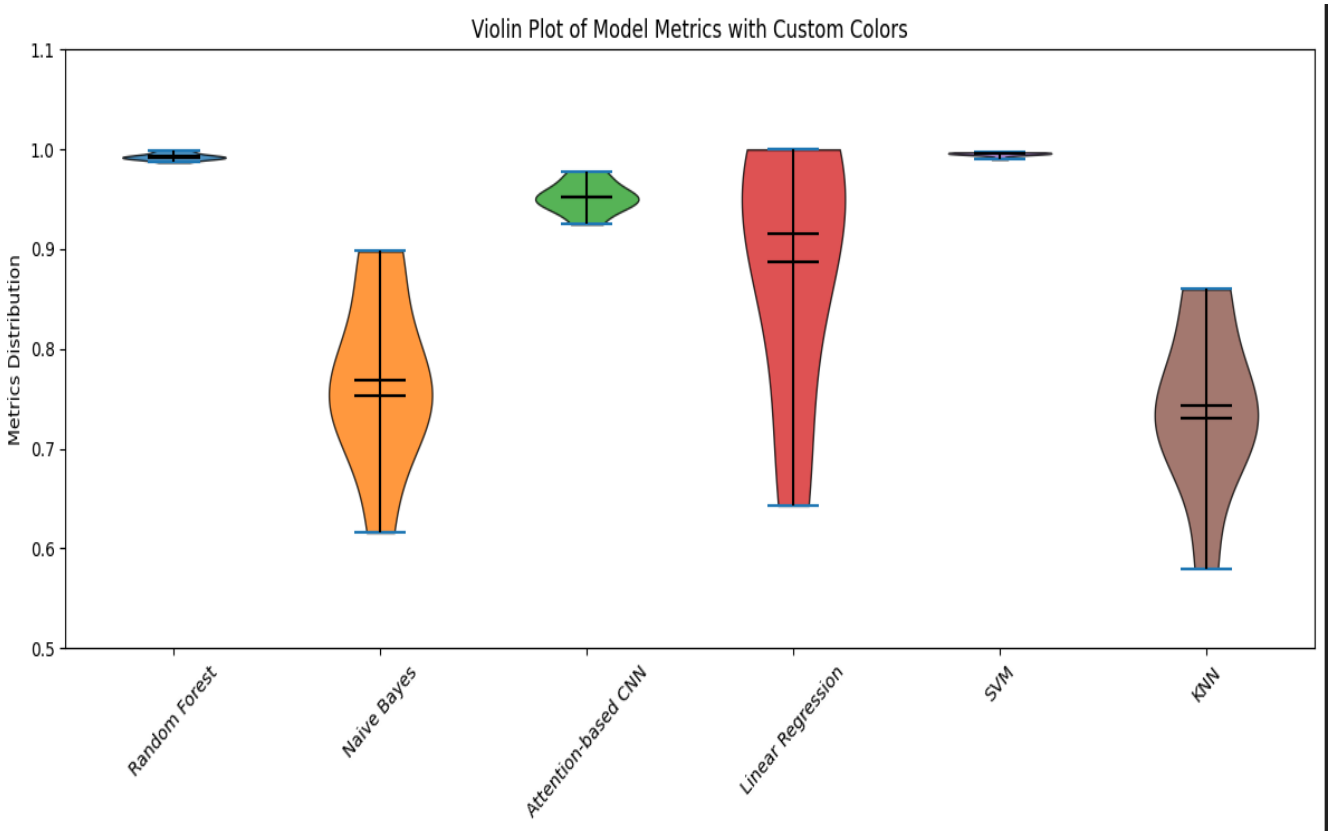


Figure 5.4: Violin Plot of Model Metrics

Figure 5.4 provides a visualization of the distribution of metrics for various machine learning models. Here, our six machine learning models are Random Forest, SVM, KNN, Attention Based CNN, Linear Regression, Naive Bayes. Metrics Distribution represents distribution of some performance metrics ranging from o.5 to 1.1. Each violin's width illustrates how densely packed the data points are at various metric values. Narrower parts signify a lower number of data points, whereas wider areas indicate a higher density.

Horizontal black lines within the violins mark the quartiles (including median and interquartile range).If present, blue vertical lines display the complete data range. So we can say that both Random Forest and SVM exhibit narrow distributions close to 1.0, reflecting stable high performance. The Attention-based CNN shows a pronounced peak in its distribution, implying

minimal variability. Naive Bayes, Linear Regression, and KNN display broader distributions, suggesting greater variability in their metrics. Linear Regression covers an extensive range, indicating inconsistent performance.

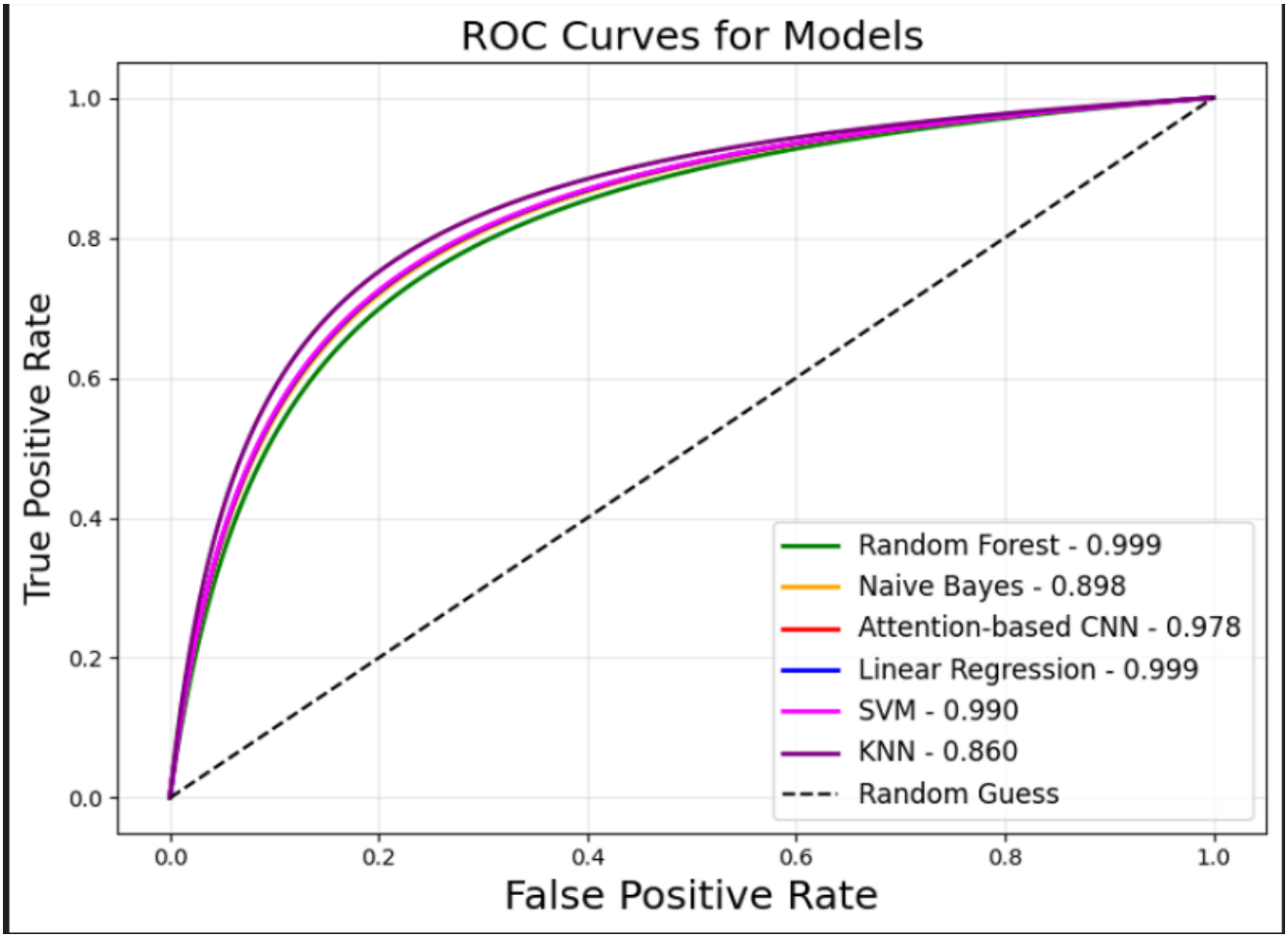


Figure 5.5: ROC Curves for Model

Figure 5.5 shows a Roc Curves for our machine learning model. In this Figure, there are two axes: one is False Positive Rate and other one is True Positive Rate .The dashed line signifies random guessing, corresponding to an area under the curve (AUC) of 0.5.

Displays the models along with their AUC scores, which measure the model's ability to differentiate between classes:

Random Forest and Linear Regression: 0.999 (High Performance).

SVM: 0.990 (High performance).

Attention-based CNN: 0.978 (Strong performance).

Naive Bayes: 0.898 (Moderate performance).

KNN: 0.860 (Satisfactory performance).

Above the Figure 5, there is no denying the fact that SVM, Random Forest, Linear Regression gives a very high performance and KNN registers the lowest AUC, yet it still outperforms random guessing. Naive Bayes also gives a well performance but Attention Based CNN model gives a strong performance that is 0.978 %. This performance is a strong and much better stable performance then other performances.

Chapter: 6 Conclusion and Future work

In this chapter, we summarize the research work and final concluding remarks with some few directions for future works.

6.1 Summary of the Research

In this research paper, we are talking about machine learning and deep learning techniques for predicting football match outcomes (English Premier League). We use six machine learning models to get better results from these models. These models were trained with data from twenty seasons to get the best model to predict football match outcomes. By examining a large set of historical match data over 20 seasons, this study reveals crucial performance indicators and external elements that affect match outcomes. Among the six models tested, the Attention-Based CNN model distinguished itself through its strong and reliable performance (0.978 %), establishing it as the best option for forecasting football match results. Other models, including Random Forest and SVM, also showcased impressive performance but fell short in terms of the flexibility and accuracy offered by the deep learning model.

6.2 Future Research Direction

Building on this work, future studies may investigate the following possibilities:

- Examination of sophisticated methods such as transfer learning and ensemble deep learning models to improve predictive capabilities.
- Expansion of the research to encompass other prominent leagues or international competitions for wider relevance.

- Integration of real-time data like player performance, weather conditions, and in-game events to enhance the accuracy of the models.
- Incorporation of interpretability frameworks to render model outputs clearer and more useful for stakeholders.

References

1) Bunker, R. P., and Thabtah, F. A machine learning framework for sport result prediction. Applied computing and informatics 15, 1(2019), 27–33.

2) Stefani, R.T.: Football and basketball predictions using least squares. IEEE Transactions on Systems, Man, and Cybernetics 7(2), 117–21 (1977)

3) Igiri, C. P., and Nwachukwu, E. O. An improved prediction system for football match results. IOSR Journal of Engineering (2014).

4) Saiedy, S., Qachmas, M., and Amanullah, F. Predicting epl football matches results using machine learning algorithms. International Journal of Engineering Applied Sciences and Technology 5 (2020), 83–91.

5) FIFA. Fifa index. https://www.fifaindex.com/ (accessed 24 October 2024).

6) Football Data. Football results, statistics & soccer betting odds data. https: //football-data.co.uk/ (accessed 24 October 2024).

7) Haghighat, M., Rastegari, H., Nourafza, N.: A review of data mining techniques for result prediction in sports. Advances in Computer Science: an International Journal 2, 7–12 (2013).
8) Bologna, C., De Rosa, A.C., De Vivo, A., Gaeta, M., Sansonetti, G., Viserta, V.: Personality-based recommendation in e-commerce. In: CEUR Workshop Proceedings. vol. 997. CEUR-WS.org, Aachen, Germany (2013).
9) Bunker, R.P., Thabtah, F.: A machine learning framework for sport result prediction. Applied Computing and Informatics 15(1), 27–33 (2019).

10) Sinha, S., Dyer, C., Gimpel, K., Smith, N.A.: Predicting the NFL using twitter. In: Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics co-located with ECML PKDD 2013. pp. 28–38 (2013).

11) Danisik, N., Lacko, P., and Farkas, M. Football match prediction using players attributes. In 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA) (2018), IEEE, pp. 201–206.

12) Hubacek, O., Sourek, G., Zelezn´y, F.: Exploiting sports-betting market using machine learning. International Journal of Forecasting 35(2), 783–796 (2019).

13) Khan, J.: Neural network prediction of NFL football games. pp. 9–15 (2003) .

14) Pappalardo, L., and Cintia, P. Quantifying the relation between performance and success in soccer. Advances in Complex Systems 21, 03n04 (2018), 1750014.

15) S. Sathe, D. Kasat, N. Kulkarni and R. Satao, “Predictive Analysis of Premier League Using Machine Learning”, I. J. Innovative Research in Computer and Communication Engineering, vol. 5, no. 3, pp. 4121-4124, 2017.

16) D. Rana and A. Vasudeva, “Premier League Match Result Prediction using Machine Learning”, Jaypee University of Information Technology, 2019.

17) Constantinou, A.C., Fenton, N.E., Neil, M.: pi-football: A bayesian network model for forecasting association football match outcomes. Knowledge-Based Systems 36, 322–339 (2012) .

18) Prasetio, D., et al. Predicting football match results with logistic regression. In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) (2016), IEEE, pp. 1–5.

19) Rodrigues, F., and Pinto, A. ˆ Prediction of football match results with machine learning. Procedia Computer Science 204 (2022), 463–470.

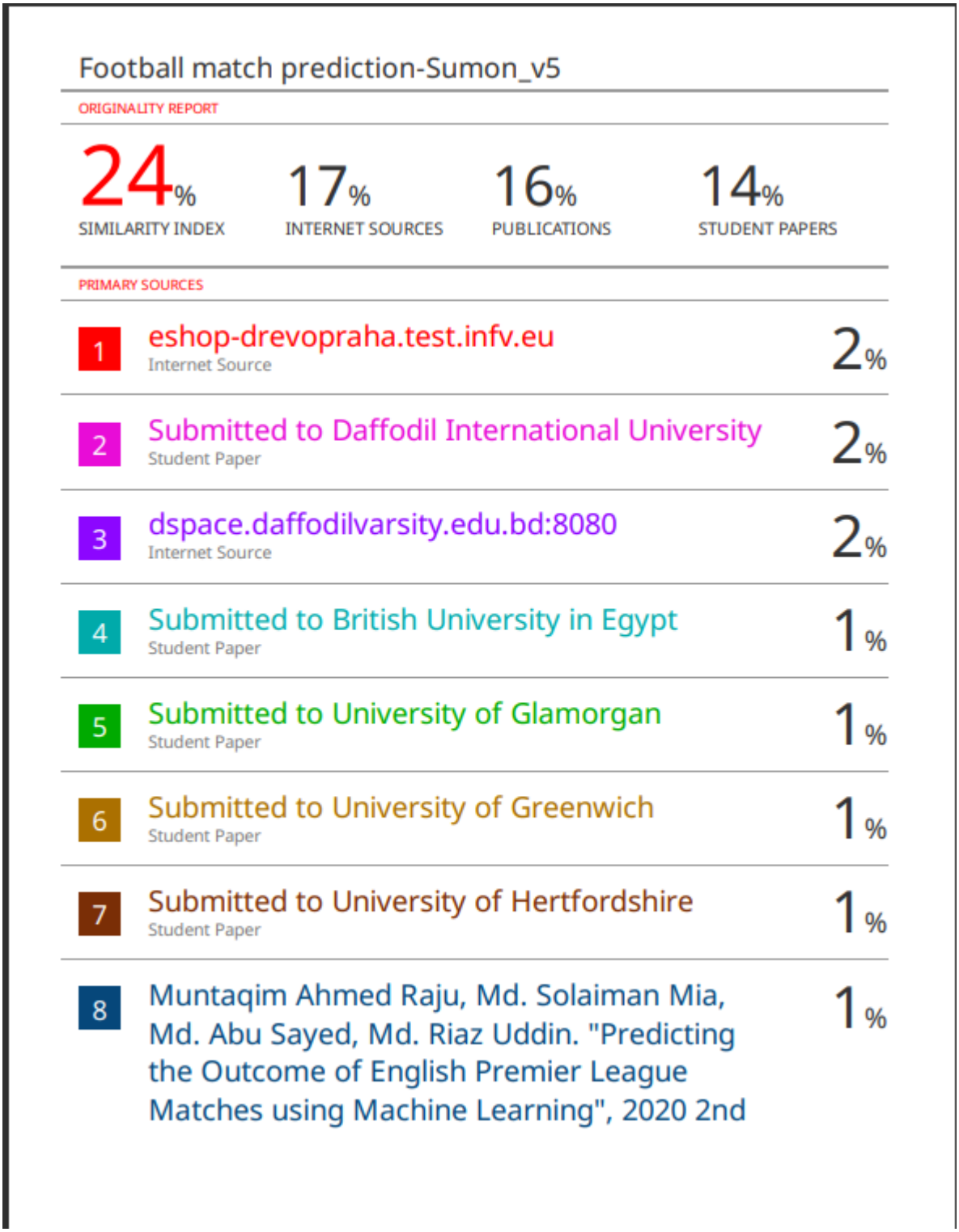
20) Baboota, R., and Kaur, H. Predictive analysis and modeling football results using machine learning approach for English premier league. International Journal of Forecasting 35, 2 (2019), 741–755.

21) Y. F. Alfredo and S. M. Isa, “Football Match Prediction with Tree Based Model Classification”, I. J. Intelligent Systems and Applications, vol. 11, no. 7, pp. 20-28, 2019.

22) Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting: Prediction of match results based on player characteristics. Applied Sciences, 10(1), 46.

23) Prasetio, D. and Harlili, D., (2016) International Conference On Advanced Informatics: Concepts, Theory And Application, pp. 1-5.

Plagiarism Report



Football match prediction-Sumon_v5

ORIGINALITY REPORT

24%	17%	16%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	eshop-drevopraha.test.infv.eu Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	Submitted to British University in Egypt Student Paper	1%
5	Submitted to University of Glamorgan Student Paper	1%
6	Submitted to University of Greenwich Student Paper	1%
7	Submitted to University of Hertfordshire Student Paper	1%
8	Muntaqim Ahmed Raju, Md. Solaiman Mia, Md. Abu Sayed, Md. Riaz Uddin. "Predicting the Outcome of English Premier League Matches using Machine Learning", 2020 2nd	1%