Recruitment of Suitable Football Player by using Machine Learning Techniques

Msc Research Project
Data Analytics

Anamika Chavan
Student ID: x18199950

School of Computing
National College of Ireland

Supervisor: Pierpaolo Dondio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Anamika Chavan
Student ID:	x18199950
Programme:	Data Analytics
Year:	2019
Module:	Msc.Research Project
Supervisor:	Pierpaolo Dondio
Submission Due Date:	12/12/2019
Project Title:	Recruitment of Suitable Football Player by using Machine Learning Techniques.
Word Count:	6652
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Recruitment of Suitable Football Player by using Machine Learning Techniques

Anamika Chavan
x18199950

Abstract

In football, player selection is the critical task which leads to a key of success. Coaches and managers form their team by selecting finest players all around the globe. Each player has his own identity in the team. When the team gets formed and due to some circumstances player moves from one club to another then finding replacement of such a player is again a big task. Finding closest match for the replaced player is the big headache that coaches and manager face. The problem illustrated in this paper is to find closest match for the replaced player by using machine learning algorithm. The players will get classified based on their ratings. In this research six machine leaning algorithms namely SVM, LDA, Naive Bayes, Decision Tree, XGBoost, KNN, have been implemented. Research compares the performances of these machine learning algorithms by using evaluation metrics such as accuracy, precision etc. LDA and SVM performed the best with accuracy of 83.77% and 80.31%. Further, KNN function will give the closest match among the predicted players.

Keywords: *Machine Learning, Prediction, Multi-classification, LDA, SVM, KNN*

1 Introduction

Football is one of the most popular sports watched and played by millions of people. It is played over 200 countries and directed by FIFA (Fédération Internationale de Football Association) (Pariath et al.; 2018). This sport is played between two teams and each team consists of 11 players. Managers and coaches are in the hunting of the most suitable players to form a noble team. The player selection process is the most important task in the pursuit of winning. Poor team selection and non-desirable combination of the players can also adversely affect on the integrity of the players Each player has an important role in the game and selecting the wrong player can cost the football team to lose the championship and even millions of the dollars. Managers and coaches select the best team players according to their requirements, wages and best prices. Player selection requires lots of effort and critical analysis. After all this critical analysis, when the team is formed and due to some circumstances when the player moves to another club. Then, finding a desirable replacement for such a player is again a big task. Thus, a prediction model that will help to decide whether a new team member is a suitable replacement for a left team member would be helpful.

1.1 Motivation and Project Background

The player recruitment process for a particular club in football intends to select the most desirable player for a particular position in the team. Each player plays an important role in the team. Managers and coaches recruit best team players for their teams with the best resources they have. Poor team selection can cause lots of problems. Thus, managers and coaches always phase the problem of player recruitment. Selecting the best player requires an analysis of qualitative and quantitative attributes. These attributes may include the player's skills, fitness, performances and anthropocentric. After completion of team player selection, when due to some circumstance's player moves from one club to another than finding the replacement of that player is again a big problem that managers face. In this case, a predictive model will provide more objective results in less amount of time.

There are several studies has been done on the prediction of a suitable player for a team. Author Mathew et al. (2018) offered the most related research in this study by proposing a predictive model for player selection. In this research, he has built different classification models and try to predict the rating of the player. Another similar kind of research is done by authorPassi and Pandey (2018) where he tries to predict the performance of the player depending upon the previous matches. For this research, he has used data like wickets taken by the bowler and run scored by a batsman. In the research of AlShboul et al. (2017), the author has used a neural network model to predict a suitable player for the team. But with this model author got a non-significant result. The researches till date have been carried out further does not suggest the best player or closet match for the replaced player.

This research implements the machine learning model which will shortlist the closest match for the replaced player by using similarity measure. Further, the result has been verified with evaluation metrics like Accuracy, Precision, Recall, and F-measure.

1.2 Research Question

This research is mainly about creating machine learning models that will classify suitable football players for a replaced player in the minimum amount of time and with good accuracy. Furthermore, closest player for the replaced players is predicted. Hence our research question is *"Can we predict suitable match for the replaced football player by using fast gradient boost technique and similarity measure?"*

1.3 Research Objectives

Obj1: A critical literature review on classification, gradient boosting algorithms and prediction of players.
Obj2: Extraction of player's data from the Kaggle.
Obj3: Implement and evaluate the result of Support Vector Machine(SVM).
Obj4: Implement and evaluate the result of Linear Discriminant Analysis(LDA).
Obj5: Implement and evaluate the result of Naïve Bayes.
Obj6: Implement and evaluate the result of Decision Tree.
Obj7: Implement and evaluate the result of K- Nearest Neighbor(KNN).
Obj8: Implement and evaluate the result of XGBoost.
Obj9: Predicting the closest player by similarity measure(Euclidean Distance).
Obj10: Comparison between the models.

The remaining sections of the paper are organized as follows. Section 2 investigates the literature review. Section 3 describes the methodology used for this research. Section 4 gives an idea about project specifications. Section 5 and Section 6 describe the implementation and evaluation strategy used for this research. Section 7 conclude and give future direction for this research.

2 Related Work

2.1 Introduction

Different researches have been carried out for the prediction of a suitable candidate for the team. The literature review section carries out the study of different research done on the classification of a suitable person for a given job using machine learning algorithms.This section is divided into four section. The first section is player prediction and classification where we have reviewed the papers which are aiming for prediction in sports industry.The second section contains intense literature review about prediction of candidate in different industries. The third section gives brief idea about studies which have achieved good results by using boosting algorithms.The last section is primarily giving idea about measure metrics in KNN.

2.1.1 Player prediction and classification

The triumph of any sports game depends upon the right player selection. There are many researchers have worked on player prediction for various sports games like football, cricket and hockey. Mathew et al. (2018) predicted suitable replacement for a football player who has transferred to another club. To accomplish this research, the authors have used different classification models. The main motive was to predict the rating field column from the dataset. If the rating of the player matches with the replaced player then that player will get selected. This study compares the performance of different machine learning models and also identifies how the number of classes affects the accuracy of the model. In the result analysis, it is found out that LDA performance was better than other algorithms.

Passi and Pandey (2018) has predicted the player's performance as how many wickets will bowler take and how many runs will batsman score. Both the problems are considered as classification problems where number of wickets and the number of runs are divided into a number of ranges.The aim of this study mainly predicting player's performance for each match. For this prediction, classification models have been used. Authors have used different attributes as no. of innings, batting average, Strike rate, etc to predict the player performance. After applying machine learning algorithms on these attributes, it is found out that decision tree and random forest worked well as compared to other algorithms.

AlShboul et al. (2017) has used a Competitive neural network model to select the players for a sports team. The main motive of this study was to select the best opponent team and predict the chances of the win. By using neural network author is predicting the rating field of each player to calculate the chances of team victory. By considering player's features separately for both the team, the neural network analyses the final result of win and loss. Firstly, the neural network has been applied to 11 players and it gave an accuracy of 54%. In the second experiment, a neural network has been applied to 22 players and it gave an accuracy of 60%.

Bunker and Thabtah (2019) has proposed a similar kind of player prediction technique by the neural network. The dataset contains different attributes such as historical performance of the teams, results of the matches and player's data which will help to predict the stakeholders the odds of winning. Attributes were grouped into four categories player resistance, player speed, physical status, and player technique. After applying neural network model accuracy was calculated. In the result analysis, it is found out that neural network predicts player but gives less accuracy.

The work carried out in the above section depicts that traditional algorithms like LDA and decision trees give a good performance. Therefore, we have implemented LDA and Decision tree models for the player selection for a football club. On the other hand neural network did not give good performance in the represented work.

2.1.2 Candidate Prediction in different industries

To reduce a large number of efforts in job recruitment, many researchers have worked on the prediction of a suitable candidate for various job roles.

e jannat et al. (2016) have applied a machine learning algorithm for candidate prediction in a software firm which will reduce manual task. For this study, the Naïve Bayes classifier was used. In this study, the dataset was ranked according to the knowledge level and parameters like GPA ranked scored in various technologies. After calculating probability and frequency, the Naive Bayes model was applied to calculate the most frequent class which will help to predict the candidate. In the result analysis, it was found out that naïve Bayes performed well on training as well as test dataset.

Jantan et al. (2010) also predicted human talent by using a decision tree classifier. The main objective of this study was to predict the employees who are deserving candidates for promotion. For this study, the author has used a decision tree C4.5 classifier. The main motive behind using decision tree classifiers as it is more favorable for a categorical outcome can deal with noisy data and easy to interpret. The features were used to predict the candidate for promotion includes skills, activities, knowledge, work outcome and contribution. In the result analysis, it was found out that the decision tree classifier worked very well with the accuracy of 95%.

A similar kind of work done in the research of Jantan et al. (2009), where authors have applied machine learning algorithms on forecasting human talent. Attributes as knowledge, management skill, individual skill, and previous performances are considered as factors for potential talent. The target classes were lecturer, senior lecturer, professor and associate professor. After the dataset pre-processing, classifier models as Random Forest, C4.5, Multilayer Perceptron, K-star and Radial Basis Function Network were applied. In the result analysis, it is found out that C4.5 gave the highest accuracy.

Apatean et al. (2017) done research to predict candidates for a specific position in a company. The main aim of this research is to help the company with a large number of CVs by passing them through the machine learning model and identify a suitable candidate for a preferred position. In this study, the position considered as a target class and candidates information considered as attributes. The dataset used for this research contains 17 classes and 18 attributes. Attributes like Education, Programming Languages, Salary, Range, etc were considered in the dataset. After the data pre-processing, data mining algorithms as KNN, LDA, Naïve Bayes and decision tree were applied on the dataset. LDA and Naïve Bayes worked better than other models.

Similar kind of research done by Author Li et al. (2011) where he created a qualitative recruitment system via SVM and MCDM approach. For this research online questionnaire test was used as dataset then SVM was used to predict the appropriate candidate while multi-criteria decision systems (MCDM) were imposed on the performance of the model. The result showed the proposed system is qualified for recruitment purposes.

From this section, we get to know that many researchers have used classification algorithms like LDA, Naive Bayes, decision tree and SVM. Therefore, LDA, Naive Bayes and Decision tree models are used for this research.

2.1.3 Classification with boosting algorithms

Classification models do not perform optimally for high dimension models. Boosting algorithms rewrite the weights based upon the previous predictor's result 8. Hence it is known as stage-wise additive modeling 9. Boosting algorithms aggregates weak classifiers into the strong classifier. Therefore, it performs better than other traditional classifiers and gives better accuracy.

Oughali et al. (2019) uses the XGBoost and Random forest algorithms to predict the shooting win of the basketball players. This study aimed to analyze the session dataset with the help of that NBA players can prepare their plans according to the other team players. After the pre-processing of the data models were applied to the dataset. In the result analysis, it is found out that XGBoost performed well than the traditional model. The main reason behind the success of XGBoost as it builds the trees sequentially, where each additive tree gives less error.

Jain and Nayyar (2018) researched employee attrition by using the XGBoost model. The main objective of this research was to predict whether the employee is leaving or staying in the organization. The IBM HR dataset was used for this analysis. According to the correlation matrix, fewer contributing attributes were pruned and the XGBoost model was applied on the dataset. XGBoost gave outstanding accuracy in the result analysis. XGBoost incorporates features such as regularization, parallel computing, flexibility and availability. Hence, Author highly suggests using the XGBoost model.

Duan and Ma (2018) had proposed a coupon usage prediction system by using the XGBoost algorithm. This research was mainly to check whether the coupons are used by users. In this study, the XGBoost model was continuously optimized by using the grid search method. Experimental results state that the XGBoost model contribute better to precision marketing. The AUC value of the model was 84%.

From this section, it can be seen that the boosting algorithm like XGBoost seems to be more accurate and faster than other boosting algorithms. Therefore, XGBoost model has been used for this research.

2.1.4 KNN

Bhannarai and Doungsa-ard (2016) used in methodology to predict the person is suitable for agile methodology or not. As agile is a popular software development tool that stresses on a collaboration of people. The agile process contains different roles as developer, tester, product owner. Each member of the team will have a different personality. One of the dangers that the manager always faces is the capability of a member to fulfill the role. By applying a personality test with classification technique project managers will know to limit their software team. This paper is predicting the agile person with KNN methodology. After pre-processing of the data author has applied KNN methodology with a different number of k. This study shows that KNN with a large number of k values gave better accuracy.

Majid et al. (2014) introduces a novel approach for colon and breast cancer prediction using different features. The research carried out in two stages: the pre-processor and the predictor. In the first stage, the mega-trend diffusion technique is used for an imbalanced dataset. The real dataset has been used in this study which includes the human protein sequence of cancer/non-cancer, breast/non breast cancer, colon/ non colon cancer. In the second stage machine learning approaches SVM and KNN are used to develop a hybrid model. The conclusion which was got from the research is that both models worked moderately in the research.

As the KNN is based on distance metric therefore classification performance can affect by distance. Author Yean et al. (2018) proposed the analysis of emotion between stroke patients and normal people. For this study, the author used the KNN classifier with different distance metrics like Euclidean, City block, Cosine, etc. Electroencephalogram (EEG) signals of various emotions are used as a dataset for this research. As the data is continuous therefore similarity metrics like Jaccard and Hamming distance does not apply to this dataset. After applying the KNN model with different distance metric author calculated the accuracy. Euclidean and city block distance measures gave better accuracy than another distance metric.

Sarma et al. (2017) came up with the idea of detection of threat with face recognition by using KNN classification. The author has proposed an authentication by verifying the facial features of the user with the addition of username and password. KNN algorithm will classify users into 4 categories such as legitimate, possibly legitimate, possibly not legitimate and not legitimate groups to verify image authenticity to detect insider threat. Smart QoS services have been imposed for better security performance. After applying the model on data, it is found that the KNN model performed well.

Author Li et al. (2017) proposed market manipulation detection with the help of supervised machine learning algorithms. The main aim of this research was to detect market manipulation in china based on the information given by data in the security market and the China Securities Regulation Commission. To achieve this goal, tick stock data and daily stock data of 64 manipulated stocks were used in this study. After data pre-processing, machine learning models were developed, such as Logistic Regression, KNN, ANN, LDA, QDA and SVM. K-fold cross-validation was applied to this model to test the robustness of the models. In the result analysis, it is found out that KNN and SVM were the most effective classification algorithms which gave over 99% accuracy.

The work carried out in above section depicts that KNN algorithm can give better performance and can be used to find the distance. As our data is mostly numeric and from the above review paper, it is found out that euclidean distance works better for numerical measure.Therefore, we have implemented KNN model with euclidean distance for the player selection for a football club.

2.2 Conclusion

Ihe work carried out in literature review section depicts that algorithms like Support vector machine, Linear Discriminant analysis, Decision Trees, Naive Bayes gave more accuracy and good performance as compare to other machine learning algorithms. In boosting algorithms, XGBoost gave the better accuracy and result. In the last section, we found out that euclidean measure works best in KNN for numerical data.

3 Methodology

3.1 Introduction

This section gives detailed idea about methodology used for this project. It also gives justification to use KDD methodology over other methodology.

3.2 Methodology

Data mining-based study requires complex processing and decision making. In this research, we have used Knowledge Discovery in Databases (KDD). According to Tso and Yau (2007), KDD describes the entire information discovery cycle and gives important insights from the data.Hence it is mostly used in the field of academic and commercial world. The KDD process is iterative and interactive in nature which invloves numerous steps where decision is made by the user. KDD mainly focuses on data mining rather than project management which makes it suitable for classification and prediction. From the below Figure 1 we can see that there are 6 phases involved in the KDD process.

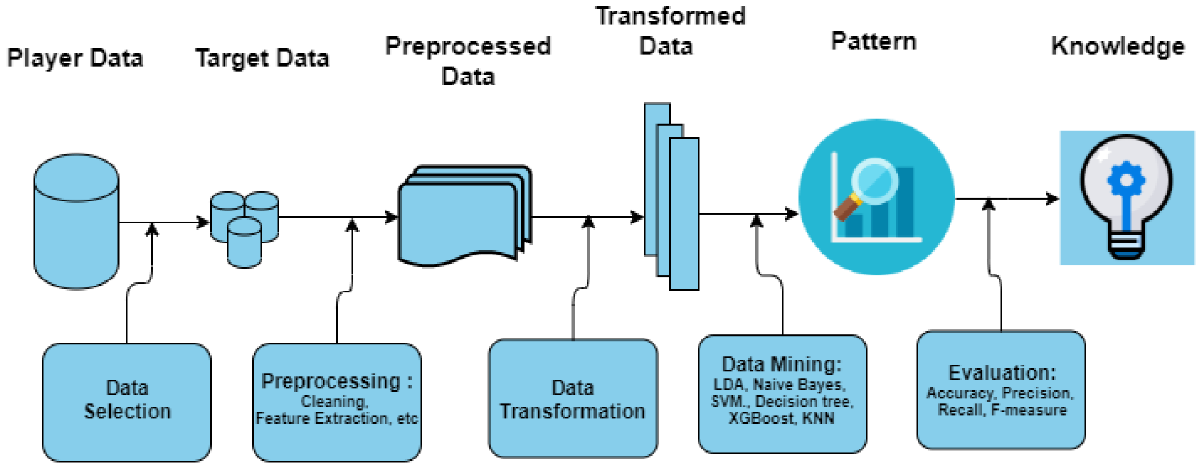
Player Data
Target Data
Preprocessed Data
Transformed Data
Pattern
Knowledge
Data Selection
Preprocessing : Cleaning, Feature Extraction, etc
Data Transformation
Data Mining: LDA, Naive Bayes, SVM., Decision tree, XGBoost, KNN
Evaluation: Accuracy, Precision, Recall, F-measure

Figure 1: Methodology for player selection

The overview of each phase is as follows.

- **Data Selection :** Data selection is the phase where task-relevant data is collected and extracted. For this research, we have collected the football player's data from the Kaggle which is available to download in CSV format. This dataset contains about 17,000 players data with 54 different attributes of the players.
- **Pre-Processing :** This phase deals with pre-processing of the data to use it for machine learning models. The data has to be pre-processed before to use on any machine learning models. If the data is not refined then accuracy of the machine learning model can get fluctuate. Usually, 80% of the efforts are usually taken by this phase. As the downloaded data contains lots of noise so it's necessary to

clean the data. Missing values check, feature Engineering, binning, etc. are a few important steps that if we missed, might bring wrong results.

- **Data Transformation :** This phase deals with finding the relevant from the data and convert it into an appropriate format before using any machine learning model. Many machine learning models are based on numerical calculations and if we feed the categorical data to these models then it will convert it into wrong format and accuracy of the model can get affected. In this phase,all categorical data converted into a numerical format. Further, scaling is applied on numerical data to ensure all features are in one unit.

- **Data Mining :** Data mining is the process where it extracts implicit and useful information from the data. It gives hidden and undiscovered information about the data.The insights discovered from data mining can be useful in fraud detection,marketing and scientific discovery.In this stage, we have created different models by using R packages like MASS and e0171. The following data mining algorithms have been used for this research.

  1. **LDA :** Linear Discriminant analysis is the dimension reduction technique which finds dissimilarities between two or more classes. It draws a separation line between classes and provides class separability. LDA is similar to logistic regression but when comes for parameter estimation it outperforms logistic regression. Therefore, it is more accurate, fast and powerful (Hastie et al.; 2004).
  2. **SVM :** SVM uses kernel trick for the transformation of the data and based on this transformation it finds out an optimal boundary between classes. The benefit of using SVM is that it can capture complex relationships between dataset attributes.
  3. **Naïve Bayes :** Naïve Bayes algorithm is basically based on Bayes theorem with a strong assumption of feature independence. It gives importance to each feature which makes it outstand all the models. It is widely used in real-world applications like fraud detection, email filtering, etc (Taheri and Mammadov; 2013).
  4. **Decision Tree :** Decision tree is the graphical representation of a result and decision which use to make that result. It is displayed in a sequential manner which makes it easier to visualize. This algorithm takes less time to construct than any algorithm (Sharma and Kumar; 2016). It shows the hidden relationship between the attributes which makes it more useful for classification and prediction.
  5. **KNN :** K-nearest neighbor algorithms use stored labeled instances to classify new data instances. It uses a similarity metric to calculate the distance between the stored instance and new instances. It separates unlabelled instances into well-defined groups

- **Evaluation :**This step involves the evaluation of machine learning models results. In this stage, the machine learning models will get apply on test dataset.It gives idea about model efficiency.To evaluate the model, we have used Accuracy, Precision, Recall and F-measure metrics. The higher the value of these metrics represents the better capability of the model.

3.3 Conclusion

From this section, we can conclude that KDD gives more attention on data mining phase and gives more meaningful insights from the data. Hence, this research has been implemented by using KDD methodology.

4 Design Specification

4.1 Introduction

This section gives idea about two architecture which has been used in this research. This section also gives justification about two-tier over three tier architecture.

4.2 Two-tier Architecture

Two tier architecture is used to design this research. As data is not stored on the database server, we are omitting the data tier layer from the design specification. Detail specification of each tier is as follows.

4.2.1 Application Tier

The application layer is also known as business logic layer because it contains core functionalities of the implementation. It contains logical functionality and control the process the data. Firstly, we will extract the data from Kaggle by using R. After that data will go through the process of pre-processing and transformation. Finally, data mining algorithms are applied on cleaned dataset.

4.2.2 Client Tier

The client layer is also called as presentation layer where we display the results. It fetches the results from application layer. After applying the models, we will display the results in graphical format by using powerBI.

4.3 Conclusion

As our research does not contains any database server. Hence, two-tier architecture has benn implemented for this research.

5 Implementation

5.1 Introduction

In this section, there is brief discussion about hardware and software specifications. A detailed description about the dataset is given in this section. Furthermore, different models used for the prediction have been discussed. We have used different types of models in order to compare the result and find out the most accurate model. Below figure shows the implementation process.

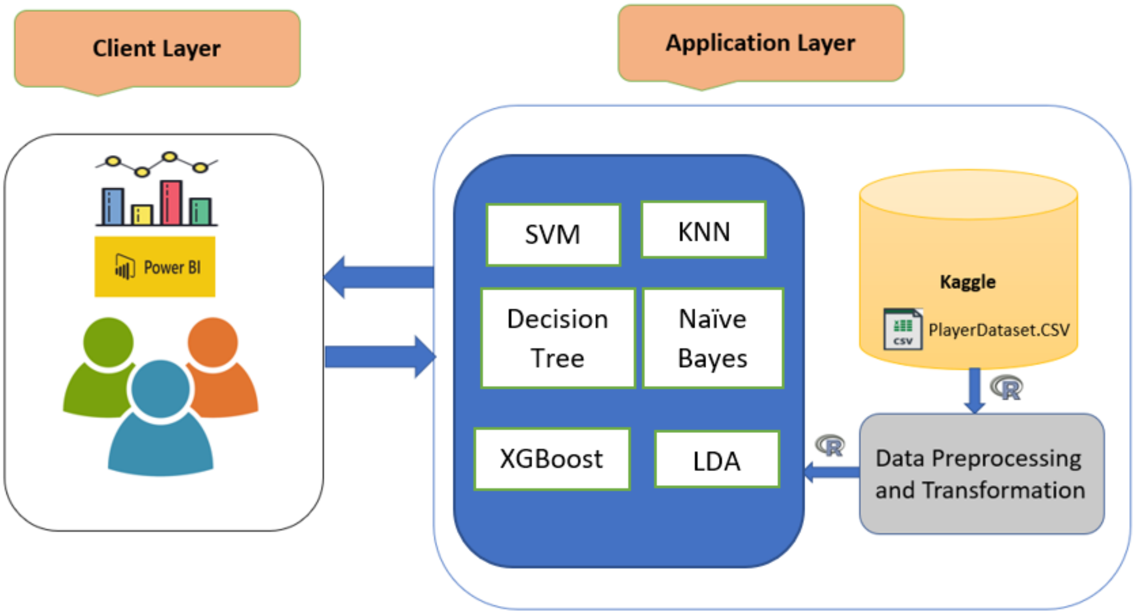

Figure 2: Two tier Architecture

5.2 Objective

The main motive of this implementation is to predict the suitable replacement for a player who moved to another club by using machine learning models. In this section, we will implement these machine learning models which will predict the rating class by using other features. Further, we will also find out the closest match for the left player.

5.3 Hardware and Software Specifications

- **Hardware Specification**
  This research have been developed in the following hardware specification environment.
  Processor: Core i5
  Operating System: Windows 10
  Storage: 256GBSSD

- **Software Specification** This whole research is implemented on R-studio. All the machine learning procedures like data cleaning, sampling and analysis are provided by R packages. To visualize the data, we have used PowerBI tool.

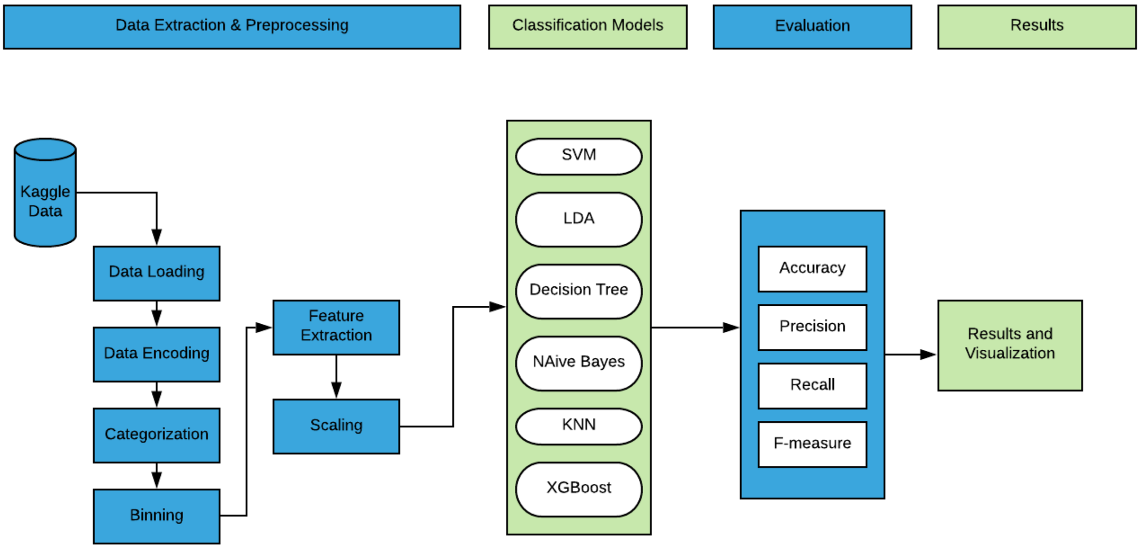

Figure 3: Implementation Flow Chart

5.4 Data Selection

The dataset is used for this project is collected from the Kaggle website [1]. It contains detail information about football players. The dataset contains information about 17,589 players and 53 attributes. The dataset contains numerical attributes as height, weight, age, vision, vision, speed, stamina, etc. and label attributes like name, club name, date of birth, etc. The dataset was downloaded into .csv format and extracted into the R studio. As the dataset is downloaded from Kaggle, we are not violating data privacy-related issues.

5.5 Data Pre-processing

To accomplish better results from machine learning models the dataset has to be in an appropriate format. Data pre-processing refers to the transformation of the raw data into clean data. If the data is not processed then it will directly affect the accuracy of the machine learning model. Hence, it is important to pre-process the data before applying it to any model. The following steps have been taken on data in order to ensure the data consistently.

1. Handling Missing Values: In this stage, we have checked for missing values in the dataset as they are a major roadblock in machine learning modeling.If we apply data without handling missing values then there are high chances of getting false output.Hence, it is important to handle missing values properly. As is.NA() function returned false for all attributes. Hence,our dataset does not contain any missing values. Therefore, we will move to a further step.

[1] `https://www.kaggle.com/hiteshp/exploring-fifa-2017-dataset`

2. Feature Engineering: In the dataset, some attributes name, club name, nationality, etc are unwanted. Retaining these unnecessary attributes will potentially blog down the runtime.Additionally, it will slow down the performance of machine learning models.Hence these attributes were removed by using R package. In this research, 10 attributes such as Name, Club, Contract_Expiry, etc were removed and machine learning models are applied on remaining 43 features.

3. Data Encoding: Machine learning models contain mathematical calculation therefore data has to be in a numerical format.If we feed raw data to machine learning model then there are chances of getting incorrect outputs.Thus, data has to be converted into numerical format before to feed any machine learning model.In this study, work_rate and preferred_foot column data are converted into the numerical format

4. Categorization of players: All football players are known for their preferred position on the ground. So, we divided all the players into 4 categories as forward, Midfielder, Defender and Goalkeeper. Hence, all machine learning models are applied on each of this categories.

5. Binning: Binning is the process of converting continuous features into categories (Bins). It reduce the effect of miner observation. Each Bin will contain a specific range of continuous values. In this study, rating field is converted in to 10 classes. The following schema is used for binning the rating filed.

Table 1: Binning in the Rating Field

Bins	**Class**	**Bins**	**Class**
40-45	1	66-70	6
46-50	2	71-75	7
51=55	3	76-80	8
56-60	4	81-90	9
61-65	5	91-100	10

6. Data Scaling: The dataset contains features that are varying in scales and units. The machine learning models which use Euclidean distance measure will be affected by this varying scale. In this, feature scaling will scale independent features in a fixed range. In this research, we have applied feature scaling on all numeric features.

5.6 Methodology

In order to develop a machine learning models for predicting a suitable football player in a replacement of a player who moved to another club, a vast variety of literature review has been done in the section 2. According to the researchers most proven machine learning algorithms which provides highest accuracy were SVM, LDA, Random Forest, Naïve Bayes, XGBoost and KNN. To assure the accuracy of the model, we have applied cross validations on the models.

5.6.1 LDA

Linear Discriminant analysis is generalized form of Fisher Linear Discriminant. It is dimension reduction technique which is use to model dissimilarities between two or more classes. LDA increases the distance between the means of the classes while reduces variation within the classes (Tae-Kyun Kim and Kittler; 2005). It provides separability by drawing line between two different classes. The main motive of LDA is to project features from higher dimension to lower dimension in order to avoid curse of dimensionality.It is implemented by using the stratified k-fold cross validation of fold 10 on the dataset to ensure that single fold is representation of overall dataset.

Further, LDA is implemented by using lda() function classifier which is downloaded from MASS package. Then this classifier is utilized on test data to predict the values of the class. Model is evaluated by using confusion matrix. In this research, the performance of the LDA model is measured by using accuracy, precision, recall and F-score.

5.6.2 Naïve Bayes

Naïve Bayes classifiers are probabilistic classifiers which are based on Bayes theorem assumes that all features are independent. In classification, for parameter estimation it requires small amount of training set (Trovato et al.; 2016). It can be trained systematically in classification depending on the nature of probalistic model.It can also applied on huge dataset as it does not consist complex repetative parameter estimation technique. It predict classes faster than any other models.

It is executed by using stratified k-fold implementation on the dataset. Then naiveBayes() is applied on the training set. This classifier is obtained from e0171 package. Then this classifier is utilized on test data to predict the values of the class. Model evaluation is done by using confusion matrix. The performance of Naïve Bayes classifier is measured by using accuracy, precision, recall and F-score.

5.6.3 Decision Tree

Decision tree constructs the classification models in the pattern of tree structure. It divides the data into smaller and smaller subgroup where the final result is a leaf node.outcome is represented by branch, attribure in the decision tree is represented by a node and leaf node contains tha class label. Construction of decision tree is always faster than other algorithms (Sharma and Kumar; 2016). When the tree consturction is completed then decision tree try to remove useless brach with the leaf node.This process is known as prunning.

It is performed by using stratified k-fold on the dataset. Then rpart () classifier is applied on the training set. This classifier is obtained from rpart package. Then this classifier is applied on test data to predict the values of the class. Model is evaluated by using confusion matrix. The performance of decision tree classifier is calculated by using accuracy, precision, recall and F-score.

5.6.4 SVM

SVM is supervised machine learning algorithm which performs classification by finding hyperplane between the classes. The maine motive of this algorithm is to find a hyperplae which will maximize the distance between the classes.It is popular in addressing multiclassification problems.

It is executed by using stratified k-fold on the dataset. Then svm() classifier is applied on the training set. This classifier is obtained from e0171 package.Then this classifier is applied on test data to predict the values of the class. Model evaluation is done by using confusion matrix. The performance of SVM classifier is measured by using accuracy, precision, recall and F-score.

5.6.5 XGBoost

XGBoost is associated with the family of boosting algorithms as it uses a gradient boosting algorithm at its core. Boosting algorithms aggregates weak classifiers into the strong classifier.XGBoost is the ensemble method of decision tree classifier.It provides faster tree prunning.

It is implemented by using stratified k-fold on the dataset. Then xgboost() classifier is applied to the training set. The XGBoost classifier is obtained from the package XGBoost. Then this classifier is applied to test data to predict the values of the class Model is evaluated by using the confusion matrix. The performance of the XGboost classifier is measured by using accuracy, precision, recall and F-score.

5.6.6 KNN

KNN is non-parametric classification method. It classifies new points based on the distance measures (Euclidean, Manhattan, etc). It gives prediction in less amount of time and there is ease to interpret those outputs.

It is implemented by using stratified k-fold on the dataset. The function that is used to implement the knn is bascially uses euclidean distance to find the closest point.Then knn () classifier is applied on the training set with the value of k=10. The knn() classifier is obtained fron the package FNN.Then this classifier is applied to test data to predict the values of the class Model is evaluated by using confusion matrix. The performance of KNN classifier is measured by using accuracy, precision, recall and F-score. Further, to calculate the closest match of the replaced player,attr() function of KNN has been used.

6 Evaluation

6.1 Introduction

There are many evaluation metrics are available to test the model. As we have applied stratified cross-validation on the dataset. Data is divided into 10 folds in which 9 folds are of training set and 1 is of test set. Validation all the models was done on the test dataset. To evaluate the models Accuracy, Precision, Recall and F1-score has been used which are given in below Table 3.

Table 2: Evaluation Table

Algorithms	Accuracy	Precision	Recall	F-measure
SVM	83.77%	0.82	0.56	0.67
LDA	80.31%	0.79	0.66	0.72
Naive Bayes	71.42%	0.58	0.54	0.56
Decision Tree	62.95%	0.64	0.42	0.51
KNN	69.52%	0.45	0.66	0.54
XGBoost	68.25%	0.56	0.58	0.57

6.2 Accuracy

Accuracy is the ratio of the number of correct predictions to the total number of predictions (Hossin and M.N; 2015). Accuracy describes weather a model is being trained correctly or not. In short, the higher the accuracy better the model. Accuracy works better for a balanced predictor class. Thus, after evaluating accuracy scores for all models, it can be seen from the table that SVM scored a high accuracy of 83.77% followed by 80.77%. It is evident from the table that the decision tree gave very little accuracy. We have calculated accuracy for different values of K which is given in the table. It can be seen that a K value decreases the accuracy of the model gets decreases. Thus, for better accuracy, we consider the value of k=10.

Table 3: Accuracy with different values of K

K-Values	Accuracy
k=3	64.47%
k=5	68.38%
k=8	68.79%
k=10	69.59%

6.3 Precision

Precision is also called as positive predicted values. It describes how the classifier correctly classified each class. It is the ratio of true positive values to the true positive and false positive. Since this study contains multi-class classification, the precision value is aggregated from each class. The precision value of the SVM classifier is highest which means SVM has a low false-positive rate. On the other hand, Naïve Bayes has a low precision value which means the model predicts more of false positive.

6.4 Recall

Recall means how many true positives are predicted. It is the ratio of true positives predicted by the model to the true positive and false negative. Again, the average value has been taken by considering the recall value for each class. Recall values for all models are plotted in the Table 3. It is evident from the graph LDA has high recall value while the decision tree has the lowest recall value.

6.5 F-measure

F-measure combination of precision and recall where it takes harmonic mean of precision and recall. F-measures gives equal weight to precision and recall. The F-score is 1 is considered a good score. From the Table 3, it is evident that LDA has a higher F-measure value followed by SVM. On the other hand, the decision tree has the lowest f-measure value.

6.6 Discussion

In this research paper, we are predicting a suitable match for a football player who moved to another club. The goal of this study is to predict the rating field of the players and find the closest match for the left player. It inputs a variety of parameters like height, age, work_rate, etc to predict the outcome variable rating. We are not only finding the rating of the player but we are also finding the closest match for the replaced player which brings novelty to this piece of work. The in-depth discussion in the literature survey in section 2 showed that SVM, Naïve Bayes, Decision tree, LDA, GBM and KNN are some consistently used algorithms in the field of sports and candidate prediction. Hence, all these models have been executed to the current research to find a suitable match for the replaced player. For the closest match of the player, Euclidean distance measure has been used which is present in the KNN model. The Figure 4 shows the comparison between evaluation metric accuracy and f-measure. From the figure, it is evident that SVM gave a good performance as compared to other models. Player prediction model by Mathew et al. (2018) achieved the highest accuracy in LDA and SVM. However, our current research also achieves the highest accuracy in SVM of 87.33% and LDA 80% which is quite commendable. A similar type of approach for candidate prediction for a specific position is done by Apatean et al. (2017) where LDA performed well with less minimal error. Also, it is seen that Jantan et al. (2010) had done human talent prediction by using the decision tree. In the result analysis of this research, it can be seen that the model worked well where our decision tree model gave an accuracy of 62.95%. The overall result of this research shows that the proposed prediction worked well.

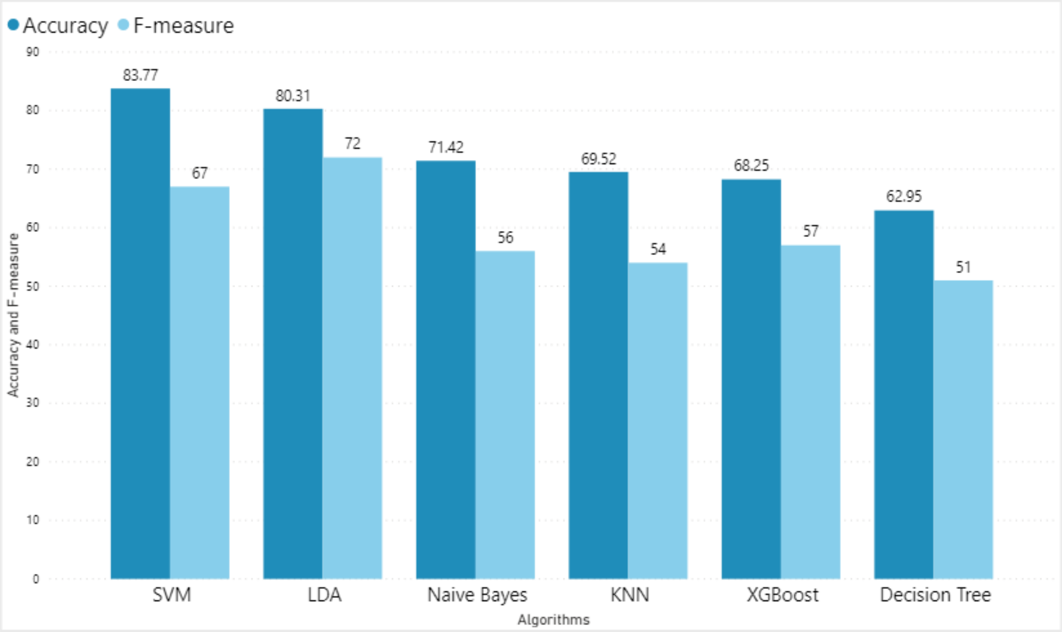

Figure 4: Accuracy and F-measure score by %

6.7 Conclusion

In this section, we applied evalutaion metrics like Accuracy, Precision, Recall and F-measure on the dataset.After that,accuracy and F-measure have been compared. Thus, we can conclude that algorthm like SVM and LDA gave good performance in terms of evaluation matrix. On the other hand, decision tree gave very less accuracy.

7 Conclusion

This research work focuses on implementing a football player prediction model for a player who moved to another club. This model will further give the closest match to the replaced player. A series of experiments have been done to find out the best model. Thus, supervised machine learning algorithms such as SVM, Naïve Bayes, LDA, Decision tree ,XGBoost and KNN were implemented to find out the best accuracy from the model. The result shows that SVM and LDA have the highest accuracy among the other classifications algorithms. The accuracy and F1- score obtained from this model are quite appreciable. To ensure the model accuracy, K-fold stratified strategy have been applied on all algorithms. KNN algorithms gave higher accuracy with the K value of 10. Further, closest player match for the replaced player have been calculated by using KNN method which uses Euclidean distance. The implemented research would be helpful to predict suitable players for a replaced player with the closest match. The main motive of this paper is to help managers and coaches to find out the best team players for their team. This solution can be extended to solve human resource problems where they need to find out a suitable candidate for various job profiles.

7.1 Future Work

In the current research,classification algorithms were used to predict the discrete nature feature. In future work, regression algorithms can also be used to predict the continuous number feature. In this way, both the features like discrete and continuous can be predicted for this field.Additionally, unsupervised methods can give a different view of the data.

8 Acknowledgment

I would like convey my appreciation to professor Pierpaolo Dondio for his valuable guidance throughout the coursework.His suggestions and advice helped me a lot during this thesis. I am also thankful to my friends, colleagues and relatives for their support and encouragment during this research.

References

AlShboul, R., Syed, T. Q., Memon, J. and Khan, F. M. (2017). Automated player selection for a sports team using competitive neural networks.

Apatean, A., Szakacs, E. and Tilca, M. (2017). Machine-learning based application for staff recruiting.

Bhannarai, R. and Doungsa-ard, C. (2016). Agile person identification through personality test and knn classification technique, *2016 2nd International Conference on Science in Information Technology (ICSITech)*, pp. 215–219.

Bunker, R. P. and Thabtah, F. (2019). A machine learning framework for sport result prediction, *Applied Computing and Informatics* **15**(1): 27 – 33.
URL: *http://www.sciencedirect.com/science/article/pii/S2210832717301485*

Duan, G. and Ma, X. (2018). A coupon usage prediction algorithm based on xgboost, pp. 178–183.

e jannat, M., Chowdhury, S. S. and Akther, M. (2016). A probabilistic machine learning approach for eligible candidate selection.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2004). The elements of statistical learning: Data mining, inference, and prediction, *Math. Intell.* **27**: 83–85.

Hossin, M. and M.N, S. (2015). A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining Knowledge Management Process* **5**: 01–11.

Jain, R. and Nayyar, A. (2018). Predicting employee attrition using xgboost machine learning approach, *2018 International Conference on System Modeling Advancement in Research Trends (SMART)* pp. 113–120.

Jantan, H., Hamdan, A. R. and Othman, Z. A. (2009). Classification techniques for talent forecasting in human resource management, *in* R. Huang, Q. Yang, J. Pei, J. Gama,

X. Meng and X. Li (eds), *Advanced Data Mining and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 496–503.

Jantan, H., Hamidah, Hamdan, A. and Othman, Z. (2010). Human talent prediction in hrm using c4.5 classification algorithm, *International Journal on Computer Science and Engineering* **2**.

Li, A., Wu, J. and Liu, Z. (2017). Market manipulation detection based on classification methods, *Procedia Computer Science* **122**: 788 – 795. 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
URL: *http://www.sciencedirect.com/science/article/pii/S1877050917326868*

Li, Y.-m., Lai, C.-y. and Kao, C.-p. (2011). Building a qualitative recruitment system via svm with mcdm approach, *Applied Intelligence* **35**(1): 75–88. Copyright - Springer Science+Business Media, LLC 2011; Last updated - 2011-06-23.
URL: *https://ezproxy.ncirl.ie/login?url=https://search.proquest.com/docview/873356714?accountid*

Majid, A., Ali, S., Iqbal, M. and Kausar, N. (2014). Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines, *Computer Methods and Programs in Biomedicine* **113**(3): 792 – 808.
URL: *http://www.sciencedirect.com/science/article/pii/S0169260714000029*

Mathew, V., Chacko, A. M. and Udhayakumar, A. (2018). Prediction of suitable human resource for replacement in skilled job positions using supervised machine learning, *2018 8th International Symposium on Embedded Computing and System Design (ISED)*, pp. 37–41.

Oughali, M. S., Bahloul, M. and El Rahman, S. A. (2019). Analysis of nba players and shot prediction using random forest and xgboost models, *2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–5.

Pariath, R., Shah, S., Surve, A. and Mittal, J. (2018). Player performance prediction in football game, pp. 1148–1153.

Passi, K. and Pandey, N. (2018). Increased prediction accuracy in the game of cricket using machine learning, *International Journal of Data Mining Knowledge Management Process* **8**: 19–36.

Sarma, M. S., Srinivas, Y., Abhiram, M., Ullala, L., Prasanthi, M. S. and Rao, J. R. (2017). Insider threat detection with face recognition and knn user classification, *2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 39–44.

Sharma, H. and Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining, *International Journal of Science and Research (IJSR)* **5**.

Tae-Kyun Kim and Kittler, J. (2005). Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(3): 318–327.

Taheri, S. and Mammadov, M. (2013). Learning the naive bayes classifier with optimization models, *International Journal of Applied Mathematics and Computer Science* **23**.

Trovato, G., Chrupała, G. and Takanishi, A. (2016). Application of the naive bayes classifier for representation and use of heterogeneous and incomplete knowledge in social robotics, *Robotics* **5**: 6.

Tso, G. and Yau, K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, *Energy* **32**: 1761–1768.

Yean, C. W., Khairunizam, W., Omar, M. I., Murugappan, M., Zheng, B. S., Bakar, S. A., Razlan, Z. M. and Ibrahim, Z. (2018). Analysis of the distance metrics of knn classifier for eeg signal in stroke patients, *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, pp. 1–4.