

# **ANALYSIS OF AUTOMATIC GENDER PREDICTION IN SOCIAL MEDIA BY USING XGBOOST ALGORITHM**

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering  
degree in Computer Science and Engineering

By

**Emmady Varsha (38110144)**

**Pasagadugula Akhila (38110388)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
SCHOOL OF COMPUTING**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119**

**MAY - 2022**



**SATHYABAMA**

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

[www.sathyabama.ac.in](http://www.sathyabama.ac.in)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **EMMADY VARSHA (38110144)**, **PASAGADUGULA AKHILA (38110388)** who have done the Project work as a team who carried out the project entitled "**ANALYSIS OF AUTOMATIC GENDER PREDICTION IN SOCIAL MEDIA BY USING XBOOST ALGORITHM**" under my our supervision from \_\_\_\_\_ to \_\_\_\_\_

**Internal Guide**

**Dr. T. PREM JACOB, M.E., Ph.D.,**

**Head of the Department**

**Dr. S. VIGNESHWARI, M.E., Ph.D.,**

**Dr. L. LAKSHMANAN, M.E., Ph.D.,**

**Submitted for Viva voce Examination held on \_\_\_\_\_**

**Internal Examiner**

**External Examiner**

## **DECLARATION**

We **Emmady Varsha** and **Pasagadugula Akhila** hereby declare that the Project Report entitled **“ANALYSIS OF AUTOMATIC GENDER PREDICTION IN SOCIAL MEDIA BY USING XGBOOST ALGORITHM”** done by us under the guidance of **Dr.T.PREM JACOB** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

**DATE:**

**PLACE:**

**SIGNATURE OF THE CANDIDATE**

## ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr.T.Sasikala M.E., Ph.D, Dean, School of Computing, Dr.S.Vigneshwari M.E., Ph.D., and Dr.L.Lakshmanan M.E., Ph.D., Heads** of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.T.Prem Jacob M.E., Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

## **ABSTRACT**

The motivation behind this project is to break down the profiles of informal organization clients to decide the gender of the profile proprietor. To this end, various AI apparatuses have been utilized and utilized in the field of portrayal from a site clarifying client choices. This study concludes that there is a plethora of features which can be mined from a Social Media profile and can be used in identifying the gender of a profile's owner. The review depends on online media and presumes that there are numerous things that can be utilized to decide the gender of the personality proprietor. Also, subsequent to inspecting online media data, studies have shown that this sexual orientation personality exercise can be performed proficiently utilizing AI procedures with 97.30%.

## TABLE OF CONTENTS

| Chapter No. | Title                                                     | Page No. |
|-------------|-----------------------------------------------------------|----------|
|             | ABSTRACT                                                  | v        |
|             | LIST OF FIGURES                                           | vii      |
|             | LIST OF ABBREVIATIONS                                     | viii     |
| 1           | INTRODUCTION                                              | 1        |
| 2           | LITERATURE SURVEY                                         | 2        |
| 3           | AIM AND SCOPE OF THE PRESENT<br>INVESTIGATION             | 6        |
| 4           | EXPERIMENTAL OR MATERIALS AND<br>METHODS; ALGORITHMS USED | 10       |
| 5           | RESULTS AND DISCUSSION,<br>PERFORMANCE ANALYSIS           | 14       |
| 6           | SUMMARY AND CONCLUSION                                    | 20       |
|             | REFERENCES                                                | 21       |

## LIST OF FIGURES

| FIGURE NO. | FIGURE NAME                        | PAGE NO. |
|------------|------------------------------------|----------|
| 1.         | ARCHITECTURE DIAGRAM               | 8        |
| 2.         | SUPPORT VECTOR MACHINE HYPER PLANE | 11       |

## LIST OF ABBREVIATIONS

| ABBREVIATION | EXPANSION                    |
|--------------|------------------------------|
| CNN          | CONVOLUTION NEURAL NETWORK   |
| SVM          | SUPPORT VECTOR MACHINE       |
| LDA          | LINEAR DISCRIMINANT ANALYSIS |
| ELM          | EXTREME LEARNING MACHINE     |
| ANN          | ARTIFICIAL NEURAL NETWORK    |
| KNN          | K-NEAREST NEIGHBOUR          |
| ID3          | ITERATIVE DICHOTOMISER 3     |

# **CHAPTER 1**

## **INTRODUCTION**

Gender identification can be very useful in many different cases and for plenty of reasons. First of all, it can be a valuable piece of information in the area of marketing. As it is known, some recommender systems depend on features such as users' age or gender in order to make proper suggestions. Undoubtedly, there are certain items that aren't aimed at both male and female users so predicting consumer's gender is the key to an astute recommendation. Sven Ewan Shepstone et al. show a new method of using automatic age and gender recognition to recommend a sequence of multimedia items to a home TV audience comprising multiple viewers. As gender identification is an aspect that concerns lots of scientists several methods have been proposed. First of all, computer vision is a way to discriminate males from females.

## **CHAPTER 2**

### **LITERATURE SURVEY**

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system. The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

#### **Title 1: Gender Prediction from Images Using Deep Learning Techniques**

**Author: Salma Fayaz Bhat, Ab Waheed Lone, Taniya Ashraf Dar**

Mechanized human age estimation by means of facial pictures is a vital and testing errand and has bunches of potential in certifiable applications, especially since the ascent of social stages and online networking. Hearty face affirmation structures are in unbelievable demand to help fight bad behavior and mental mistreatment. Distinctive applications join giving customer affirmation to get the opportunity to control to physical and virtual spaces to ensure higher security. In any case the issue of recognizing a man by taking a data stand up to picture and organizing with the known face pictures in a

database is now an amazingly troublesome issue. This is a result of the variability of human faces under different operational circumstance conditions. For instance edification, insurgencies, appearances, camera see centers, developing, beauty care products, and eyeglasses. In this paper, a significant increase in accuracy of Gender prediction can be obtained through the use of convolution neural networks (CNN) for extracting features. Making the use of Convolution neural network (CNN) encompassed with deep learning methods, state-of the art performance has been achieved. The image-based Gender estimation is determined by performing extensive experiments on the largest public available datasets of face images with Gender labels -IMDB-WIKI dataset.

## **Title 2: Human Gender Classification using Machine Learning**

**Author: Miss. Vaishnavi Y. Mali, Dr. Babasaheb G. Patil**

Human Gender classification is one of the most interested and critical area of research. Research contains interactions between computers and human which includes vast information concerning difference in characteristics of males and females. In several kind of pattern recognition, machine learning gives a relation between gender and face. This paper proposes comparison between different techniques used for gender classification. Face is a unique biometric feature of the individual. Facial images with different combinations including frontal, aligned, smiling, non-smiling as well as expression images make the system complicated. Various face recognition methods such as Convolutional neural networks, Delaunay triangulations, geometry based methods like SVM (Support vector machine), LDA (Linear discriminant analysis).For human gender classification, SVM provides better accuracy as compared with existing methods.

## **Title 3: The construction of Chinese microblog gender-specific thesauruses and user gender classification**

**Author: Zhiliang Zhu, Zejun Ke, Jiayin Cui, Hai Yu and Guoqi Liu**

Based on the statistical features, short text messages published by different gender users are different in terms of the words and semantics used. In this paper, two new features are constructed after constructing a gender-specific thesaurus. A new classification model is constructed by combining the traditional statistical features and the improved text implicitness feature. The experimental evaluation performed on the Sina Weibo dataset demonstrated the effectiveness of gender-specific thesaurus-based features, and the improved text implicitness feature improved the accuracy of gender classification to 84.7%.

#### **Title 4: Age and Gender Prediction From Face Images Using Attentional Convolutional Network**

**Author: Amirali Abdolrashidi, Mehdi Minaei, Elham Azimi, Shervin Minaee**

In this work, we propose a deep learning framework, based on the ensemble of attentional and residual convolutional networks, to predict gender and age group of facial images with high accuracy rate. Using attention mechanism enables our model to focus on the important and informative parts of the face, which can help it to make a more accurate prediction. We train our model in a multi-task learning fashion, and augment the feature embedding of the age classifier, with the predicted gender, and show that doing so can further increase the accuracy of age prediction. Our model is trained on a popular face age and gender dataset, and achieved promising results. Through visualization of the attention maps of the train model, we show that our model has learned to become sensitive to the right regions of the face.

#### **Title 5: A hybrid deep learning CNN–ELM for age and gender classification**

**Author: MingxingDuan' KenliLi' CanqunYang' KeqinLi**

Automatic age and gender classification has been widely used in a large amount of applications, particularly in human-computer interaction, biometrics, visual surveillance, electronic customer, and commercial applications. In this paper, we introduce a hybrid

structure which includes Convolutional Neural Network (CNN) and Extreme Learning Machine (ELM), and integrates the synergy of two classifiers to deal with age and gender classification. The hybrid architecture makes the most of their advantages: CNN is used to extract the features from the input images while ELM classifies the intermediate results.

#### **Title 6: Gender Prediction Methods Based on First Names with genderizer**

**Author: Kamil Wais**

In this paper, the author recommends a set of tools (including one dedicated to gender prediction, the R package called genderizeR), data sources (including the genderize.io API), and metrics that could be fully reproduced and tested in order to choose the optimal approach suitable for different gender analyses.

#### **Title 7: Age Group Classification using Convolutional Neural Network (CNN)**

**Author: Muhammad Firdaus Mustapha, Nur Maisarah Mohamad, Ghazali Osman and Siti Haslini Ab Hamid**

Age group classification is a complex task that is used to classify facial images or videos into predetermined age categories. The proposed age group classification model achieved 84.90% accuracy for the training images and 85.12% accuracy for the test images. The experimental results showed that CNN is capable of achieving competitive classification accuracy throughout two age groups in the AAF dataset with unbalanced data distribution.

#### **Title 8: Automated Detection of Gender from Face Images**

**Author: Revathi Ramachandran Nair, Reshma Madhavankutty, Dr. Shikha Nema**

The project identifies or detects the gender from the given face images. The tools used involve Convolutional Neural Network along with programming language like Python.

The project has been motivated by problems like lack of security, frauds, child molestation, robbery, criminal identification.

## **CHAPTER 3**

### **AIM AND SCOPE OF THE PRESENT INVESTIGATION**

#### **EXISTING SYSTEM**

We compare our methods with existing machine learning methods applied to gender identification including ANN and C4.5. We use the female ratios of each community together with some explicit user profile attributes on Sina Weibo as features. We adopted 10-fold cross validation on our 1,039 samples and show the identification results of the proposed and existing methods, where the best performers of proposed methods/existing supervised methods/text-based approaches are highlighted as bold underline, respectively. From the results. Existing works on gender identification focus on selecting text-based features manually and applying machine learning algorithms on sample data collection to train classifiers.

#### **DISADVANTAGES OF EXISTING SYSTEM**

- Limited understanding of the machinery of classifiers (Black Box)
- Requires significant amount of data
- May not work in cases where data collection is difficult or expensive
- Problem of over-fitting if model fitted on small dataset.

#### **PROPOSED SYSTEM**

The proposed methodology with the main function of each processing step is as follows:

**1. Data Collection:** In this step, a great amount of Social Media profiles with all the useful information were chosen and studied.

**2. Feature Extraction:** This step includes the procedure of extracting the features that seem to come in handy in discriminating men from women.

**3. Feature Selection:** In this part, a group of the features which seem to be more representative is selected.

**4. Machine Learning Model:** Here, a machine learning model is applied to identify the gender of a profile owner, based on the features of the preceding steps.

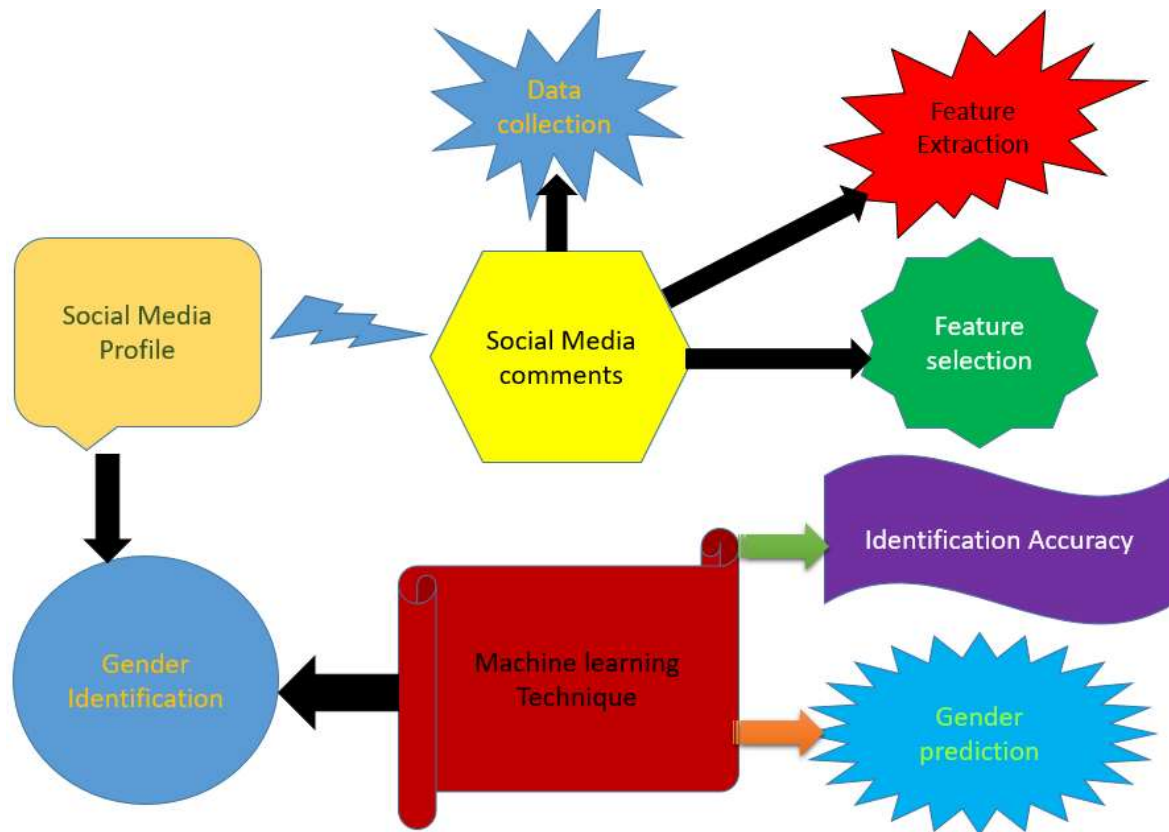
The experimental results justified the initial theoretical claims, since the gender of the Social Media profiles owners' was identified by a significant accuracy level up to 97.30%.

The proposed methodology was high accurate, there is still room for further improvement of the identification accuracy. This improvement can be achieved by collecting more data from more Social Media profiles using an automated procedure and by applying deep learning models which are able to tackle more data.

## **ADVANTAGES OF PROPOSED SYSTEM**

- Large scale deployments of Machine Learning beneficial in terms of improved speed and accuracy.
- Understands non-linearity in the data and generates a function mapping input to output (Supervised Learning).
- Recommended for solving classification and regression problems.

## ARCHITECTURE DIAGRAM



## MODULES

- Data Collection Module
- Feature Extraction and Feature Selection
- Apply Machine Learning Technique
- Gender Prediction Module

## MODULE DESCRIPTION

**Data Collection Module:**

In this step, a great amount of Social Media profiles with all the useful information were chosen and studied.

### **Feature Extraction Module:**

This step includes the procedure of extracting the features that seem to come in handy in discriminating men from women. In this part, a group of the features which seem to be more representative is selected. The feature extraction method which is used to reduce the number of resources by extracting the features of the preprocessed information for classification. It is the process of dimensionality reduction which converts the raw information into manageable groups. Face feature is a technology used in a variety of applications that classifies human faces using digital images. Machine learning is an application of artificial intelligence which provides capability to learn automatically and boost from experience without human assistance and adjust correspondingly. Learning includes observation or patterns present in the data and make well decision in future. Machine learning performance in a known task improves with the experience.

### **Apply Machine Learning Technique:**

Here, a machine learning model is applied to identify the gender of a profile owner, based on the features of the preceding steps.

- The experimental results justified the initial theoretical claims, since the gender of the Social Media profiles owner' was identified by a significant accuracy level up to 97.30%.
- The proposed methodology was high accurate, there is still room for further improvement of the identification accuracy. This improvement can be achieved by collecting more data from more Social Media profiles using an automated procedure and by applying deep learning models which are able to tackle more data.

### **Gender Prediction Module:**

The purpose of this paper is to analyze Social Media users' profile aiming at identifying the gender of the profile's owner. To this end several machine learning models were adopted and applied on a representative set of features extracted from Social Media profiles describing users' preferences relative to their gender information.

## **CHAPTER 4**

### **EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHMS USED**

#### **PROPOSED ALGORITHMS**

##### **1. XGBOOST ALGORITHM**

- XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
- In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.
- However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.
- Please see the chart below for the evolution of tree-based algorithms over the years.

##### **The algorithm differentiates itself in the following ways:**

- A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
- Portability: Runs smoothly on Windows, Linux, and OS X.
- Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.

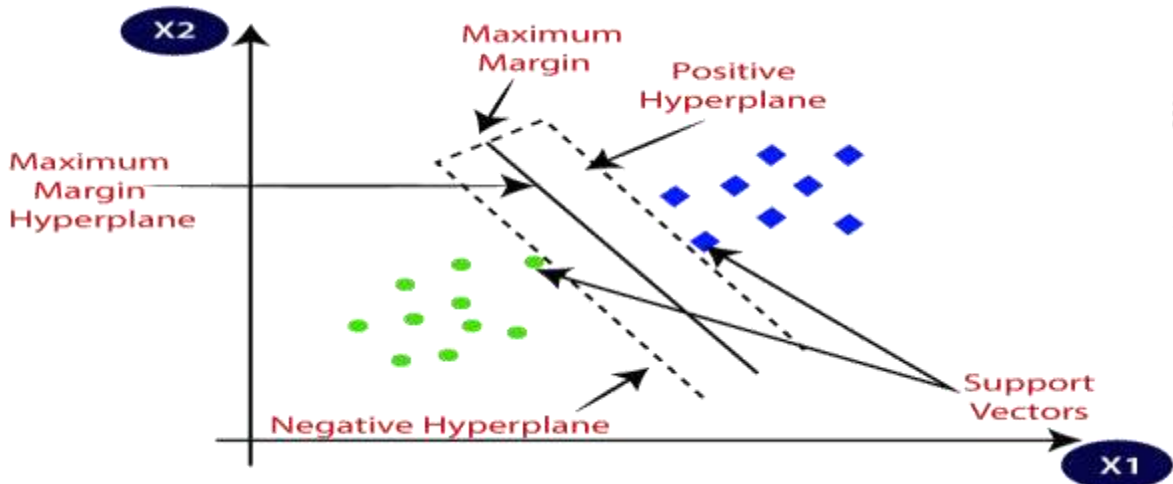
- Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

**The implementation of the algorithm was engineered for efficiency of compute time and memory resources.**

- A design goal was to make the best use of available resources to train the model.
- Some key algorithm implementation features include:
- **Sparse Aware** implementation with automatic handling of missing data values.
- **Block Structure** to support the parallelization of tree construction.
- **Continued Training** so that you can further boost an already fitted model on new data.

## **2. SUPPORT VECTOR MACHINE ALGORITHM**

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- This best decision boundary is called a hyperplane.



### Steps for Implementation of Support Vector Machine

1. Data Pre-processing step
2. Fitting the SVM classifier to the training set
3. Predicting the test set result
4. Creating the confusion matrix
5. Visualizing the training set result
6. Visualizing the test set result

### 3. Decision Trees for Classification: A Machine Learning Algorithm

- The choice tree is a sort of control machine where the data is continually circulated by specific rules (this implies the info and result of the preparation). A tree can be characterized by two elements: independent direction and wickedness. Leaves are the ultimate choice or arrangement. Choices are made based on data sharing. An illustration of a tree arrangement can be clarified utilizing a parallel tree above. Assume you need to anticipate an individual's life dependent on data like age, dietary patterns, and actual work. The themes covered are "How old are you?", "Do you work out?", "Do you work out, etc. eat a great deal of pizza? What's more, leaves are the "right" or "wrong" arrangement.

- For this situation, it was a two-section issue (type no issue). There are two primary kinds of choice trees: Sorting trees (Yes/No Type) What we saw above is an illustration of a tree, and the outcomes were variable as 'proper' or 'unseemly'. Here the choices change by class. Compromise Tree (Media Type Continued) Here the choice or reaction changes Continued, for example, Numbers like 123, etc. Movement Now that we have recognized the choice tree, we will check out how it functions inside. Numerous calculations make an answer; however, the best is known as the ID3 calculation. ID3 implies dichotomy or causes 3. Before we talk about the ID3 calculation, we will take a gander at a couple of subtleties.

#### **4. Naive bayes algorithm**

- This is an arranging strategy dependent on Base's hypothesis, which expects the freedom of speculations. More or less, Naive Bayes records it as one of the classifications that don't has anything to do with different highlights. For instance, apples are a red organic product, around 3 creeps in distance across. Albeit these attributes are because of some characteristic, these qualities assume a part in making it feasible for these seeds to be apples, consequently the name "innocent." The Naive Bayes model is not difficult to construct and exceptionally helpful, particularly for huge items. Notwithstanding its straightforwardness, Innocent Bays is known for its prevalence over the most grounded method of positioning.

#### **5. K-nearest neighbor (KNN) classification method**

K-NN is a model dependent on discovering that carries execution nearer to the nearby level and defers counting until it is executed. The k-NN calculation is the least difficult in AI calculations. Neighbours from the line of thing classifications (fork-NN classes) or property things an incentive (fork-NN in reverse) are known.

STEP 1: Start

STEP 2: Input:  $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$

STEP 3: another instance of requesting  $x = (x_1 \dots x_n)$

STEP 4: Calculate  $(x_i, c_i)$  d  $(x_i, x)$  for each case.

STEP 5: d  $(x_i), x)$  from base to top,  $(l = 1 \dots N)$

STEP 6: Select the K cases close to x:  $D_{kx}$

STEP 7: The most widely recognized classifications in  $D_{kx}$  x

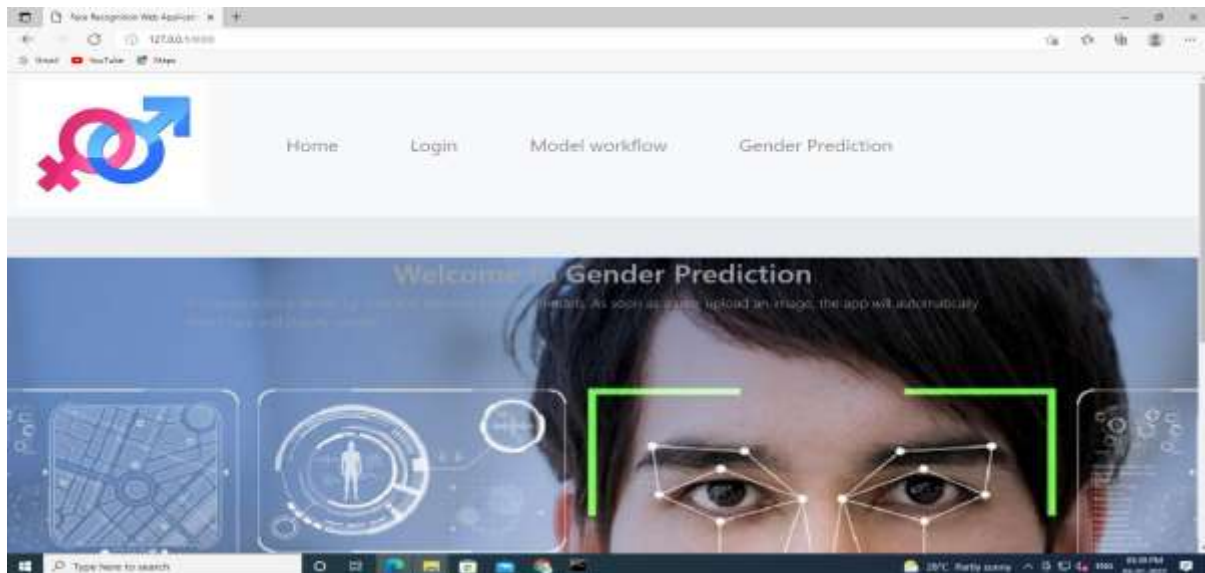
STEP 8: Return the predicted classification.

## CHAPTER 5

### RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

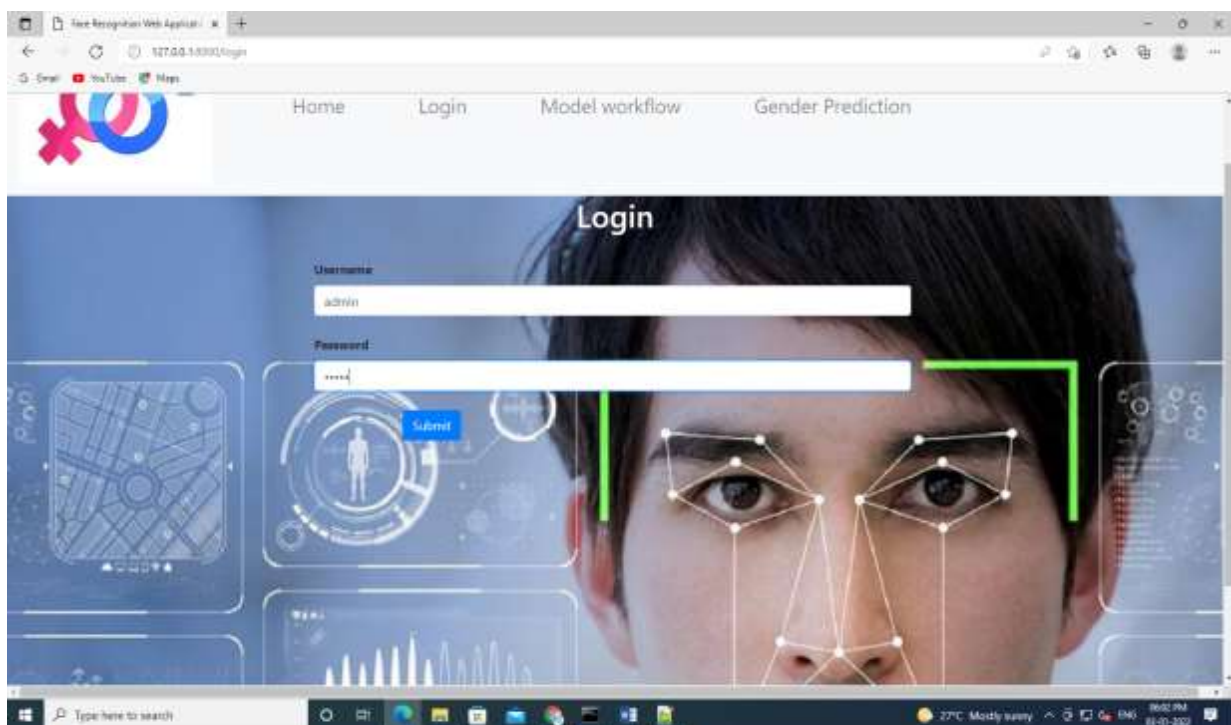
#### SCREENSHOTS

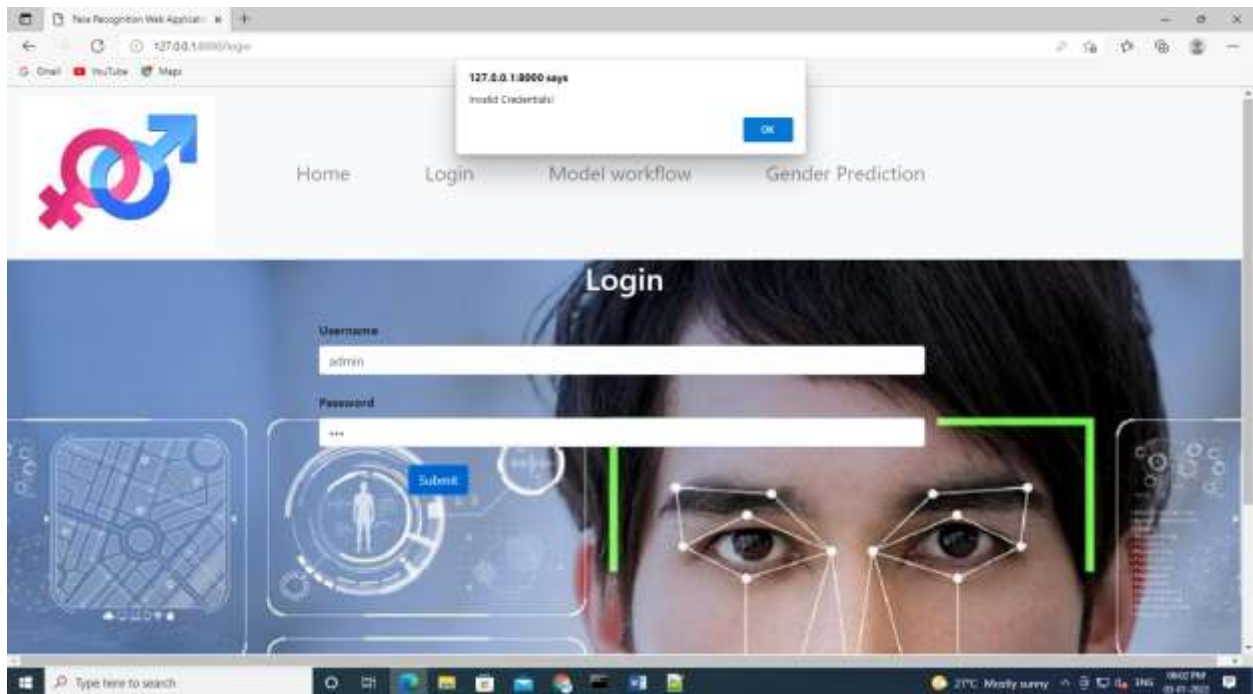
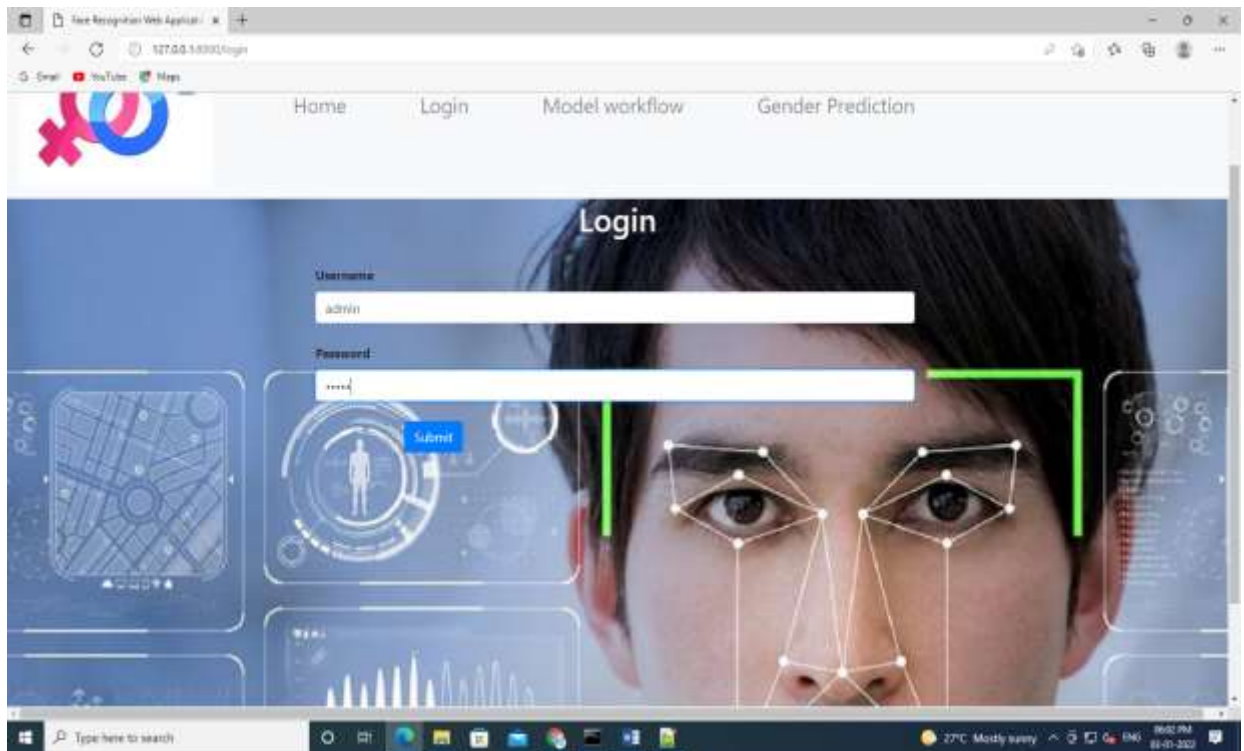
##### HOME PAGE



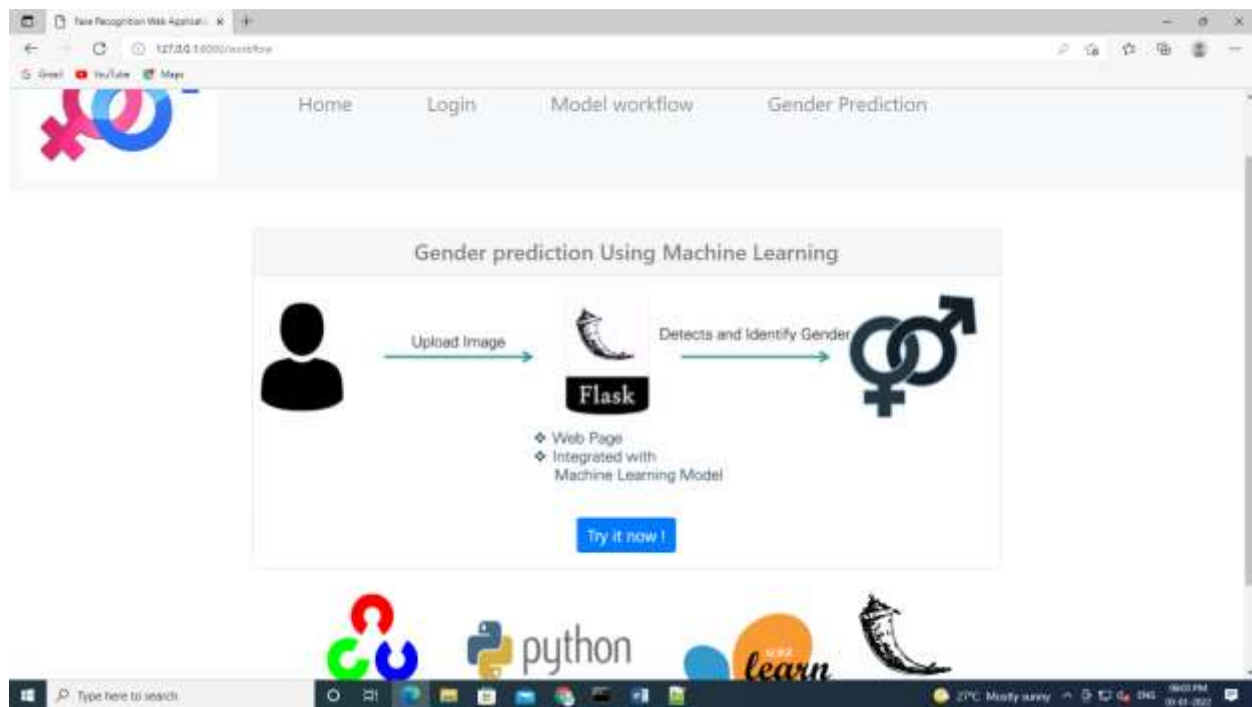


## LOGIN PAGE

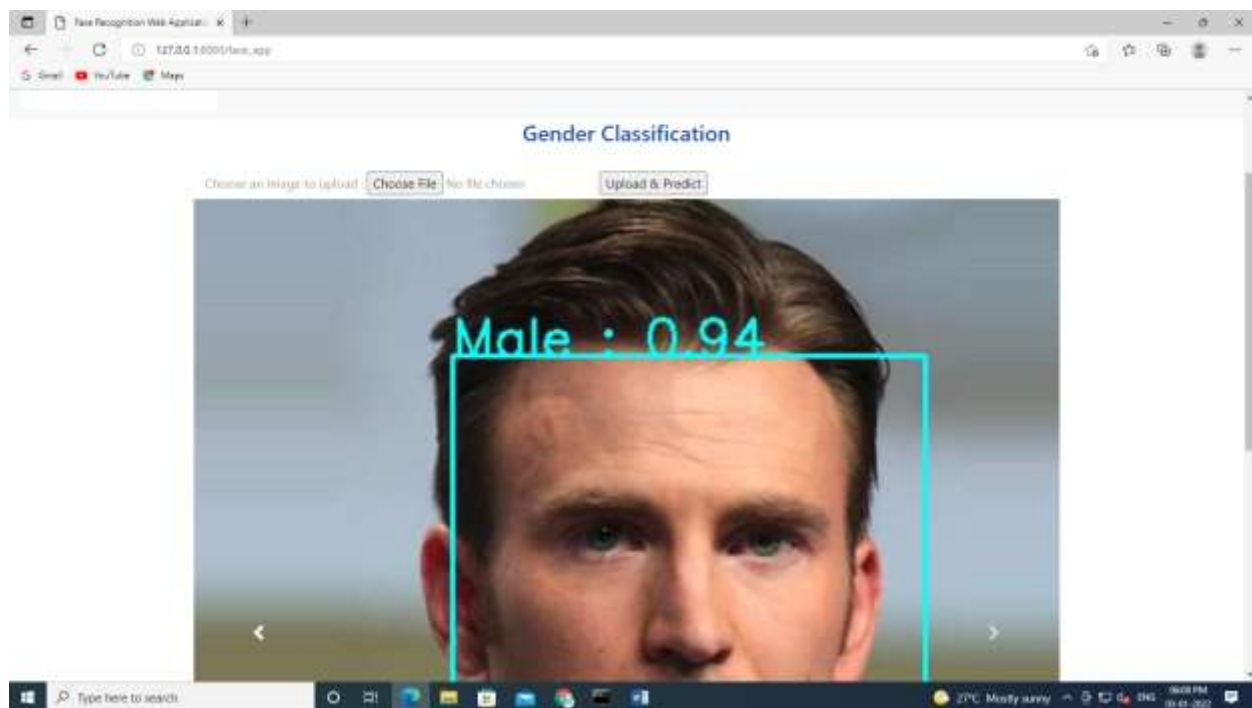




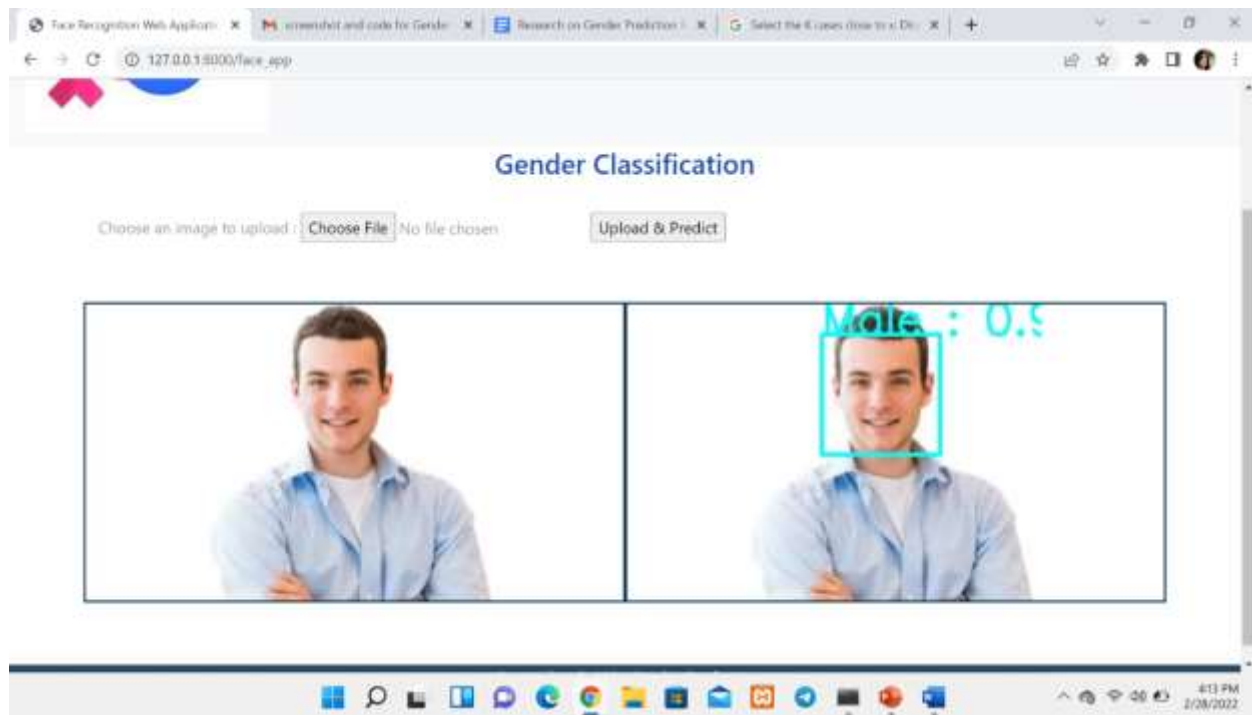
## WORKFLOW PAGE



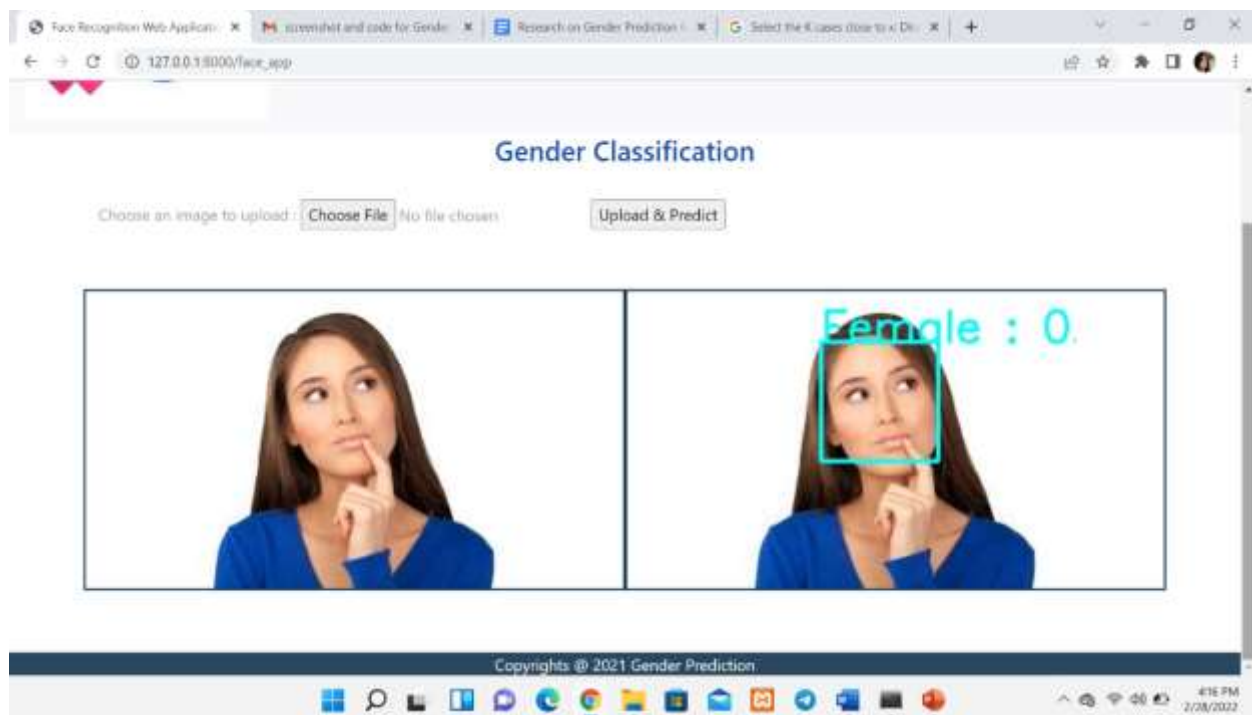
## GENDER PREDICTION PAGE



## RESULT



## PREDICTING GENDER FEMALE



## PREDICTING GENDER MALE



## PERFORMANCE ANALYSIS AND DISCUSSION

- The experimental results justified the initial theoretical claims, since the gender of the Social Media profiles owners' was identified by a significant accuracy level up to 97.30%.
- The proposed methodology was high accurate, there is still room for further improvement of the identification accuracy. This improvement can be achieved by collecting more data from more Social Media profiles using an automated procedure and by applying deep learning models which are able to tackle more data.

## **CHAPTER 6**

### **SUMMARY AND CONCLUSION**

The examination has become one of the primary hindrances to estimating interestingly, the principles are not generally something similar thought of. This prompts mistaken correlations and subsequently offers less. Studies like gender in various dialects have made around 70-80% progress. It is likewise conceivable to make more precise estimations with more data.

In a real-world, human gender classification using machine learning is broadly accepted. This paper provides detailed review of a system for human gender classification using machine learning. This system tested with only different standard dataset. The system has successfully recognized classification with good accuracy and found less accurate in some algorithms. The misclassification of the similar features can be improved using system parameters with high recognition rate.

## REFERENCES

- [1]. R Ghosh and M Dekhil, Mashups for semantic user profiles[C], Beijing, China:ACM, 2008.
- [2]. Y Zhao, Dong et al., "User Identification Based on Multiple Attribute Decision Making in Social Networks[J]", *China Communications*, vol. 10, no. 12, pp. 37-49, 2013.
- [3]. A Khan, S Jamwal and M Sepehri, "Applying Data Mining to Customer Churn Prediction in an Internet Service Provider", *International Journal of Computer Applications*, vol. 9, no. 7, pp. 8-14, 2010.
- [4]. V Eyharabide and A Amandi, "Ontology-based user profile learning[J]", *Applied Intelligence*, vol. 36, no. 4, pp. 857-869, 2012.
- [5]. B Dai, S Li, Z Gong et al., "Semi-supervised gender classification with multiple type of text[J]", *Journal of Shanxi University (Natural Science Edition)*, vol. 40, no. 1, pp. 14-20, 2017.

- [6]. S Li, J Wang, G Zhou et al., "Interactive gender inference with integer linear programming[C]", *International Joint Conference on Artificial Intelligence*, pp. 2341-2347, 2015.
- [7]. P Zhang et al., "A Gender Classification Method for Chinese Microblog Users Fused with Two Classifiers [J]", *Computer Engineering and Design*, vol. 40, no. 01, pp. 268-272, 2019.
- [8]. Y Cao, Research and Application on Gender Classification of Microblog Users.[J], Anhui University, 2019.
- [9]. Cong Liang, Shuyin Xia and Zizhong Chen, "Improvement k—nearest neighbor classification algorithm based on reference points[J]", *Computer Engineering*, vol. 45, no. 2, pp. 167-172, 2019.
- [10]. L Xu, C Jiang, Y Ren et al., "Micro-blog Dimensionality Reduction—A Deep Learning Approach[J]", *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. 7, pp. 1779-1789, 2016.