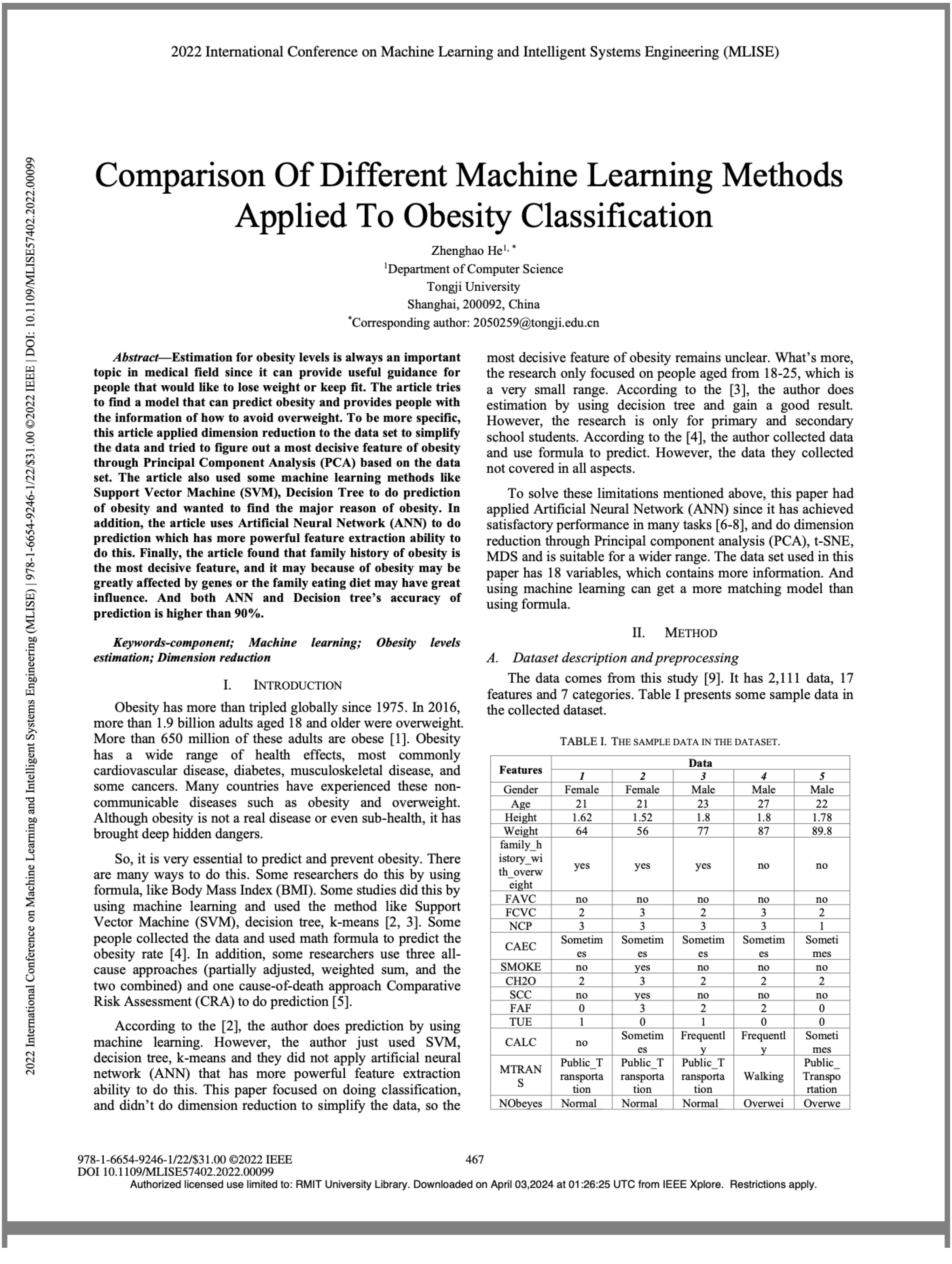

2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)

2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) | 978-1-6654-9246-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/MLISE57402.2022.00099

Comparison Of Different Machine Learning Methods Applied To Obesity Classification

Zhenghao He[1, *]
[1]Department of Computer Science
Tongji University
Shanghai, 200092, China
[*]Corresponding author: 2050259@tongji.edu.cn

***Abstract*—Estimation for obesity levels is always an important topic in medical field since it can provide useful guidance for people that would like to lose weight or keep fit. The article tries to find a model that can predict obesity and provides people with the information of how to avoid overweight. To be more specific, this article applied dimension reduction to the data set to simplify the data and tried to figure out a most decisive feature of obesity through Principal Component Analysis (PCA) based on the data set. The article also used some machine learning methods like Support Vector Machine (SVM), Decision Tree to do prediction of obesity and wanted to find the major reason of obesity. In addition, the article uses Artificial Neural Network (ANN) to do prediction which has more powerful feature extraction ability to do this. Finally, the article found that family history of obesity is the most decisive feature, and it may because of obesity may be greatly affected by genes or the family eating diet may have great influence. And both ANN and Decision tree's accuracy of prediction is higher than 90%.**

Keywords-component; Machine learning; Obesity levels estimation; Dimension reduction

I. Introduction

Obesity has more than tripled globally since 1975. In 2016, more than 1.9 billion adults aged 18 and older were overweight. More than 650 million of these adults are obese [1]. Obesity has a wide range of health effects, most commonly cardiovascular disease, diabetes, musculoskeletal disease, and some cancers. Many countries have experienced these non-communicable diseases such as obesity and overweight. Although obesity is not a real disease or even sub-health, it has brought deep hidden dangers.

So, it is very essential to predict and prevent obesity. There are many ways to do this. Some researchers do this by using formula, like Body Mass Index (BMI). Some studies did this by using machine learning and used the method like Support Vector Machine (SVM), decision tree, k-means [2, 3]. Some people collected the data and used math formula to predict the obesity rate [4]. In addition, some researchers use three all-cause approaches (partially adjusted, weighted sum, and the two combined) and one cause-of-death approach Comparative Risk Assessment (CRA) to do prediction [5].

According to the [2], the author does prediction by using machine learning. However, the author just used SVM, decision tree, k-means and they did not apply artificial neural network (ANN) that has more powerful feature extraction ability to do this. This paper focused on doing classification, and didn't do dimension reduction to simplify the data, so the most decisive feature of obesity remains unclear. What's more, the research only focused on people aged from 18-25, which is a very small range. According to the [3], the author does estimation by using decision tree and gain a good result. However, the research is only for primary and secondary school students. According to the [4], the author collected data and use formula to predict. However, the data they collected not covered in all aspects.

To solve these limitations mentioned above, this paper had applied Artificial Neural Network (ANN) since it has achieved satisfactory performance in many tasks [6-8], and do dimension reduction through Principal component analysis (PCA), t-SNE, MDS and is suitable for a wider range. The data set used in this paper has 18 variables, which contains more information. And using machine learning can get a more matching model than using formula.

II. Method

A. Dataset description and preprocessing

The data comes from this study [9]. It has 2,111 data, 17 features and 7 categories. Table I presents some sample data in the collected dataset.

TABLE I. The sample data in the dataset.

Features	**Data**				
	1	***2***	***3***	***4***	***5***
Gender	Female	Female	Male	Male	Male
Age	21	21	23	27	22
Height	1.62	1.52	1.8	1.8	1.78
Weight	64	56	77	87	89.8
family_history_with_overweight	yes	yes	yes	no	no
FAVC	no	no	no	no	no
FCVC	2	3	2	3	2
NCP	3	3	3	3	1
CAEC	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes
SMOKE	no	yes	no	no	no
CH2O	2	3	2	2	2
SCC	no	yes	no	no	no
FAF	0	3	2	2	0
TUE	1	0	1	0	0
CALC	no	Sometimes	Frequently	Frequently	Sometimes
MTRANS	Public_Transportation	Public_Transportation	Public_Transportation	Walking	Public_Transportation
NObeyes	Normal_	Normal_	Normal_	Overwei	Overwe

978-1-6654-9246-1/22/$31.00 ©2022 IEEE
DOI 10.1109/MLISE57402.2022.00099