## Introduction

Down syndrome, also known as trisomy 21, is a genetic disorder caused by the presence of all or part of a third copy of chromosome 21. The physical abnormalities that together lead to the consequence that the distinctive facial appearance associated with this condition include slanting palpebral fissures with inner epicanthic folds, flatness of the bridge of the nose, midfacial hypoplasia, and a tendency to protrude the tongue, especially when very young [1]. Down syndrome (DS) is the most frequent cause of human congenital mental retardation, which affects approximately 1 in 700 live births [2]. All people with DS exhibit mild to severe abnormalities in language and cognitive development, memory and learning impairments, and behavioral alterations [3]. Although it has been over 110 years since the first description of DS by Langdon Down, and 60 years since Jérôme Lejeune found that an extra chromosome 21 was present in individuals with DS [4]. The developmental aberrations and genetic mechanisms of various often debilitating diseases experienced by DS patients are still challenges for the society[5]. Despite many years of research, Down's syndrome can’t be cured toward current therapies [6].

Nowadays, there are plenty of researches about Down’s syndrome, appealing people to put away their prejudice and give “Down baby” a normal life. It works, Down’s syndrome is getting lot more attention. Since Down syndrome is difficult to cure, it is very important to prevent it in advance. Studying these disease-causing genes can help people better understand the control of genes on traits, so that people can better prevent disease [7]. Through the detection of gene sequences, people can know in advance the probability and risk of certain diseases in the future, and then based on the results of the detection, can they take corresponding preventive measures. For example, if one person has this disease-causing gene, he has better take medicine to stop the time of disease. Moreover, studying the effects of genes on diseases can also assist scientists that they will be more convenient to research methods to treat the disease, such as targeted therapies [8]. If scientists can further study the effects of drugs which can cure Down syndrome, scientists can use this technology to understand the different effects of the same drugs on different groups of people, so as to assist clinicians in choosing drugs that are more suitable for the patient's body. When this technology gradually matures and is widely used, it will also promote the accuracy and humanization of medical diagnosis.

Machine learning (ML) encompasses a broad range of algorithms and modeling tools, which used for a vast array of data analyzing tasks in the medical field in recent years[9]. For example, scientists had used machine learning to find that Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline[10]. [Qian Zhao](https://www.spiedigitallibrary.org/profile/Qian.Zhao-2862) and [Kenneth Rosenbaum](https://www.spiedigitallibrary.org/profile/notfound?author=Kenneth_Rosenbaum) propose a novel method to detect Down syndrome using photography for computer-assisted image-based facial dysmorphology by machine learning[11].

In this paper, the research analyzes the expression level of gene protein in mice with Down syndrome, based on logistic regression method to find out the proteins with greater influence, and to provide possible approaches for future treatment for Down syndrome. Logistic regression is the method which is used to analyze the data. Logistic regression analysis is a statistical technique which can evaluate the relationship between various predictor variables (either categorical or continuous) and an outcome which is binary (dichotomous). Aiming at the problem of Down syndrome in mice, we use logistic regression to verify and test on the UCI data set. Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters. By using this method, the research model was evaluated and Its accuracy were 94.8%, which means the method has certain validity.

Logistic Regression

Logistic regression is a common statistical analysis model, suitable for modeling discrete data. Its basic function is to perform binary classification and predict its possible labels for different input features. Logistic regression measures the relationship between a categorical dependent random variable (label) and one or more independent variables (features) by using a logistic regression function to estimate probability. Binary classification is a common task in machine learning. Linear classification performs well when the data is “linearly separable”. However, when we try to use the least square (LS) approach to fit parameter , because the features are imbalanced, this may cause the decision boundary to be pulled to a certain side of the data and lead to misclassification. Under normal circumstances, given a score z, calculate the probability that the corresponding output label is 0 or 1, the equations are as follows:

Where ,

As z increases, it is more likely to be judged as class “1”, otherwise it tends to be “0”.

From this case, corresponding to a series of discrete data points, the most suitable logistic regression model is needed to make relatively accurate predictions. Unlike linear regression using LS to find the most suitable linear model, in logistic regression, it’s able to fit the parameters of the training data by using the maximum likelihood (ML) estimation, when , the ML weights for binary logistic regression are:

Which is a common alternative expression, it also known as binary cross-entropy loss.

Mice Protein Expression Data Set was found in the UCI Machine Learning Repository. This data set includes the expression levels of 77 proteins that produce detectable signals in the cortex nucleus. There were 38 control mice, 34 trisomy mice (Down’s syndrome), 72 in total. In the experiment, each sample/mouse has 15 measurement records for each protein. Therefore, for control mice, there are 570 measurements, and for trisomy mice, there are 510 measurements. The data set contains 1080 measurement values ​​for each protein, which are all independent of each other [12]. The original data is divided into 8 classes according to genotype, behavior, and treatment. This study only focuses on the relationship between genotype and protein expression level, so it is simplified to a binary classification of 2 classes.

According to the summary information of the website, there are missing values ​​in the data set. After re-retrieving the data set, the position of the missing value is located. In order to ensure the integrity of the data and facilitate subsequent research, a new data set was created in which the missing values ​​in each column were filled with the average of the non-missing values. At the same time, the strings of “Control” and “Ts65Dn” in the “Genotype” column are converted into numbers “0” and “1”, so that logistic regression can be applied more intuitively. The last step in the data preprocessing is to adjust the 77 columns of protein expression level data to a normal distribution, which is to avoid overfitting of the data caused by the excessive variance of some features.

In order to better reflect the relationship between data volume and accuracy, we randomly sampled the data and compared the data performance at 20%, 40%, 60%, 80%, and 100%. The results are shown in the table below.

Table 1. Data Performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fraction | Precision | Recall | F1-score | Error rate  (1 - Accuracy) |
| 20% | 0.9361 | 0.9011 | 0.9135 | 0.0840 |
| 40% | 0.9405 | 0.9360 | 0.9361 | 0.0625 |
| 60% | 0.9409 | 0.9532 | 0.9467 | 0.0495 |
| 80% | 0.9508 | 0.9525 | 0.9514 | 0.0451 |
| 100% | 0.9542 | 0.9575 | 0.9554 | 0.0417 |

It can be seen that as the amount of data increases, data performance shows a steady upward trend, and the ability to capture the expression level of the protein with the highest correlation is also improved. For the training results under the complete data set (that is, the fraction is 100%), the finally drawn coefficient stem and leaf map is shown in Figure 1, and the two genes with the highest correlation are ITSN1\_N and BRAF\_N.

图表

描述已自动生成

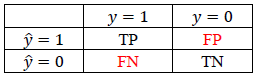
Figure 1 Coefficient Stem Plot of Logistic Regression Model

**Cross Validation**

In binary classification, there are 4 kinds of situation, including 2 types of correct: TP (True positive, the actual sample is positive, the predicted result is positive as well) and TN (True negative, the actual sample is negative the predicted result is negative as well) and 2 types of error: FP (False Positive, the actual sample is negative, but the predicted result is positive), and FN (False negative, the actual sample is positive, but the prediction result is negative).

Table1. Confusion matrix

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |



Based on the 4 situations above, 4 common machine learning performance metrics can be calculated: Precision, Recall, F1-score, and Accuracy.

Precision: The ratio of the amount of predicted true positive samples in the total amount of positive samples.

upload_805981835

The formula for precision

②

Recall: To show how many positive samples are correctly predicted.

upload_199062992

The formula for recall

③

Accuracy: The ratio of the amount of predicted true samples in the total amount of samples.

upload_377896834

④

The formula for accuracy

F1-score: The harmonic average between precision and recall.

upload_038211823

The formula for F1-score

⑤

In order to get the above consequences toward this research, the method uses 10-fold cross validation which divides mice protein expression data set into 10 subsets, then takes turns to treat 1 subset as test set and the rest of 9 subsets as train set. So it comes in for 10 models where 10 consequences can be calculated from. Averaging those consequences as the final consequence. 10-fold cross validation method uses limited data set efficiency, it is a good model to evaluate the predictive performance of the model, reduce overfitting to some extent.

The precision result is: 0.9482, the recall result is 0.9560, the error rate is 0.0463, and the f1-score result is 0.9513. So the research model is a comparatively accurate prediction model.

Table 2： title

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

不同数据量下的性能对比

Table 3： title

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Discussion

In this study, logistic regression analysis was used to study the proteins most related to Down’s syndrome in mice, and to find the corresponding gene for subsequent research on the cure of the disease.

However, there are still some areas that can be improved in the work of this research, such as the small amount of data. To solve the problem of low data volume, we have two solutions. The first is to expand the existing data volume and analyze a large amount of data to get a more accurate result. Another solution is to make full use of existing data. This study uses binary classification and only analyzes two proteins. But in fact, the table provides experimental data for eight kinds of proteins, which can be used for multivariate data analysis to enrich the experimental results. If the database is difficult to find more, there is a method named few- shot learning, which could help to solve this problem. It is well known that humans are very good at recognizing a new object through a very small number of samples. For example, children only need some pictures in the book to recognize what is "zebra" and what is "rhino". Inspired by the rapid learning ability of humans, researchers hope that after the machine learning model has learned a large amount of data in a certain category, for a new category, it only needs a small number of samples to learn quickly. This is what few-shot learning wants to solve. Applying few-shot learning to this research can also enhance the persuasiveness of the experimental results

Moreover, there are still some unsolved problems in machine learning in the field of medical research. As we all know, life is priceless, and a little problem in medicine may lead to the death of a patient. Therefore, before every new medicine or treatment is applied to the human body, they need to go through a lot of tests. However, the application of machine learning in the medical field cannot be fully explained. With the results obtained by machine learning research, people can only know that it is effective in this scenario, but the underlying principles and whether other scenarios are feasible have not been resolved yet.

Although machine learning cannot work in a leading position in the medical field, it can still be used as an auxiliary means to assist medical staff in their work. For example, for the X-ray or other medical images taken by the patient, the computer can perform simple analysis on the image through machine learning, and provide auxiliary suggestions to the doctor and the patient. Another auxiliary method is the ability of machine learning to process large amounts of data to help doctors analyze the patient's medical history. These are all medical fields that can be used in machine learning at this stage.

**Conclusion**

## References

[1] Epstein, Charles J. "Down syndrome." Abnormal States of Brain and Mind. Birkhäuser, Boston, MA, 1989. 43-44.

[2] Parker SE, Mai CT, Canfield MA et al. Updated National Birth Prevalence estimates for selected birth defects in the United States, 2004–2006. Birth Defects Res A Clin Mol Teratol 2010; 88: 1008–1016.

[3] Briggs, James A., et al. "Integration‐free induced pluripotent stem cells model genetic and neural developmental features of down syndrome etiology." Stem cells 31.3 (2013): 467-478.

[4] Lejeune J, Turpin R, Gautier M. Mongolism; a chromosomal disease (trisomy). Bull Acad Natl Med 1959; 143: 256–265.

[5] Roper RJ, Reeves RH. Understanding the basis for Down syndrome phenotypes. Plos Genet 2006; 2: e50.

[6] Mégarbané, André, et al. "The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of Down syndrome." Genetics in Medicine 11.9 (2009): 611-616.

[7] Rongvaux, Anthony, et al. "Human hemato-lymphoid system mice: current use and future potential for medicine." *Annual review of immunology* 31 (2013): 635-674.

[8] Karczewski, Konrad J., and Michael P. Snyder. "Integrative omics for health and disease." Nature Reviews Genetics 19.5 (2018): 299.

[9] Carleo, Giuseppe, et al. "Machine learning and the physical sciences." Reviews of Modern Physics 91.4 (2019): 045002.

[10] Cole, James H., et al. "Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline." *Neurobiology of aging* 56 (2017): 41-49.

[11] Zhao, Qian, et al. "Down syndrome detection from facial photographs using machine learning techniques." *Medical Imaging 2013: Computer-Aided Diagnosis*. Vol. 8670. International Society for Optics and Photonics, 2013.

[12] “Mice Protein Expression Data Set”, UCI Machine Learning Repository.