# Risk assessment using machine learning in Duchenne muscular dystrophy

***Boonsit Yimwadsana***

Faculty of Information and Communication Technology Mahidol University

***Abstract***

Risk assessment is an essential component of prognosis and treatment for genetic diseases. In the past, Mendelian inheritance analysis plays a key role in genetic risk assessment. Recently, machine learning has been widely used in data analysis with notable success. However, most of the machine learning techniques do not allow physicians to understand how data features are related to each other because the data collected by doctors are often unbalanced or bias towards patient data only. People who do not have a specific disease (healthy people or people with different diseases who may have similar symptoms), were often not followed up by physicians. Due to the data unbalance, the results of the risk assessment analysis performed by machine learning techniques are often not effective in the real-world situation. This work aims to introduce additional pedigree data to improve the accuracy of the genetic risk assessment. We tested our concept with Duchenne muscular dystrophy (DMD) disease and show that our proposed use of pedigree information help improve the accuracy even in the situation of the unbalanced data.

***Keywords:*** Sentiment analysis, Health, Natural Language Processing

*Received: 10 June 2023, Revised: 25 July 2023, Accepted: 1 September 2023*

## *INTRODUCTION*

The invention of new technologies allows most diseases to be treatable. At present, genetic diseases are still difficult to be treated using conventional methods. Complex diseases that affect systematic operation and proteins that regulate functions of human body are often caused by mutation in the genetic information of the patient. Due to the lack of treatment for genetic diseases, avoiding the disease from happening in subsequent generations may be the best option of eradicating genetic diseases. In order to avoid the diseases in subsequent generation, hosts that have high risk of passing down the genetic mutation to their offspring must be identified and consulted.

Since each human generation is rather long (approximately 30 years before giving birth to an offspring), the study of human genetic diseases often lacks sufficient data. Thanks to Mendelian theory and the discovery of human chromosome and DNA, human pedigree or family tree with genetic information can be roughly estimated.

This work focuses on a genetic disease called Duchenne Muscular Dystrophy (DMD) which is a rare muscle disorder. It affects approximately 1 in 3,500 male births worldwide. There is currently no effective cure for the disease. It is usually recognized between three and six years of age of the patients. DMD is characterized by the weakness of the muscles of the patients. As the disease progresses, muscle weakness and atrophy spread throughout the body. Serious life-threatening complications usually developed when the muscles that are responsible for the operation of the heart and respiratory system are affected.

DMD is caused by the mutation of the DMD gene on the X chromosome. The gene regulates the production of a protein called dystrophin which is thought to play an important role in maintaining the strength of the membrane (sarcolemma) of muscle cells[1]. The mutation of the DMD gene causes the diseases due to the stop in the

**Correspondence:** Boonsit Yimwadsana, Faculty of Information and Communication Technology Mahidol University, 999 Phutthamonthon Sai 4 Rd, Salaya, Phutthamonthon District, Nakhon Pathom 73170, E-mail: boonsit.yim@mahidol.ac.th

production of the dystrophin protein. Since the DMD gene occurs on the X chromosome, the disease is therefore more common in males rather than females because females carry two X chromosomes and it is rare for a female to carry two DMD-mutated chromosomes (one from father and one from mother). Even though DMD is a genetic disease, sometimes people who do not have a family history of DMD can get the disease when their genes become defective on their own due to environmental factors.

Although the gene that causes the DMD disease is well-known, how much the gene affects different patients or family is still unknown. In order to assess the relationship between the severity of the DMD gene and the DMD disease, further data analysis is needed. This works will explore popular data analysis techniques based on machine learning and determine the most suitable data analysis technique that can provide risk assessment of the disease.

## BACKGROUND

### A. Classical Risk Analysis of DMD

Mendelian theory is often applied to the study of risk analysis of genetic diseases. The pedigree of the family of the patient who has a genetic disease must be determined. In the case of DMD, we are interested in the situation that a mother is a carrier and a father is a healthy person, the risk for a female offspring to have a mutated gene from the X chromosome is generally 1/4, the chance for a female offspring to be healthy is 1/4, the risk of a male offspring to have the X chromosome is 1/4 and the chance for a male offspring to be healthy is 1/4. It is very rare for both parents to carry mutated DMD gene in the X chromosome. This is because most male died since a very young age from the disease. According to this information, we can expect that the mother of a DMD patient should be a carrier with high probability.

### B. Using Machine Learning for Risk Assessment of Genetic Disease

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

The field of machine learning has progressed significantly and there are a lot of applications of machine learning for prediction and classification problems. Risk assessment of genetic disease can be viewed as a classification problem[3, 4]. Common classification methods such as Random Forest[6], Support Vector Machine (SVM) using radial and linear kernels[7], Gradient Boosting [8], and Neural Networks[8] are often used to determine the best model. Usually the process of predictive modeling for risk assessment include 1) data exploration to determine the quality of the data and the obvious predictor variable or feature that is highly associated with the disease 2) dealing with data imbalances 3) data splitting 4) model learning and model selection 5) model evaluation and prediction and 6) variable importance. The accuracies for each technique were found to be quite high (usually above 0.7) and the SVM technique is often found to be performing the worst. After the best model is selected after using cross-validation technique, the variable importance (important features) is determined. It is often the case that clinically proven affecting genes would have a very high association with the disease.

## METHODS

Our work focuses on the risk assessment of patients who have DMD disease and the risk assessment of the relatives of the target patients in Thailand. The patient data were collected after the patients gave consent to participate in this research upon the doctor visit. The data collected from the patients include well-known disease indicator from medical literatures including DMD gene and Creatine Kinase (CK) blood test (which is a significant factor determining the DMD disease) in addition to patients' profiles and DMD disease indicators (e.g., the brain functions including motor and memory). In Duchenne, a lack of dystrophin leads to the breakdown of muscles in the body from tolerating the constant muscle movement of everyday activities. This gives rise to tiny tears in the damaged muscle membrane. The CK enzyme leaks out of the muscle through these tears and into the blood. DMD genetic test called MLPA test is performed as well as genetic sequencing of patient blood. Other vital blood test information such as sugar level, liver enzymes, and counts of red blood cells and white blood cells were also collected. In order to determine severity of the disease, the ability to walk, heart and lung function, levels of sleepiness and concentration, and biopsies of muscles may be performed in order to validate the low level of dystrophin in the muscle.

In our research approximately 143 patients including their relatives around 232 people) were studied. We will

use Bayesian analysis technique in order to assess the risk associated with the disease according to the data we collected because Bayesian analysis technique allows us to perform classification with output in the form of probability. In a sense, Bayesian analysis can be considered as a machine learning technique which incorporated the probabilistic analysis of the Mendelian analysis together.

We used machine learning algorithms that follow the standard data science and scientific methodology involving the design–learn–test processes. Standard classification techniques were used in order to identify the best model that can predict the occurrence of the disease with some knowledge about important data features which affect our prediction of the disease. Our goal is that a machine learning algorithm can predict the occurrence of the DMD disease with accuracy at least 80 percent and the best model can recognize certain features that are important to the prediction of the disease.

### *A. Data Exploration*

We explore the data that we collected from the physicians and found that most physicians collect data supporting the traditional Mendelian analysis especially pedigree. Because of this, all patient data contains individual family trees in addition to the indicator and disease severity data that we aim to collect. Figure 1 shows an example of a patient family tree which follows typical x-linked recessive inheritance. The data that we collected are mostly biased towards patients who have developed the symptoms of DMD and people who are carriers of DMD (female with one mutated DMD gene in the X chromosome).

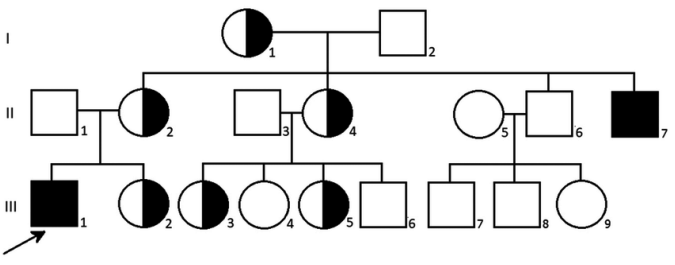


Figure 1: an example of a family tree for a DMD patient. For explanation of standard family tree symbols, see[2].

### *B. Data Balancing*

It is typical to see that the physicians do not record any data for the relatives in the same family tree as the patients. So we can assume that the healthy relatives have around average or normal measurements for blood test and genetic test. Using this information, we can simulate data for healthy people and data set become more balanced.

### *C. Data Cleansing, Transforming, and Splitting*

Once the data set is cleaned, balanced, and ready to be analyzed, some data such as the family tree data which is usually in the form of a drawing will have to be transformed into text data which can be analyzed by Bayesian analysis and machine learning techniques[10]. Each patient represented as a node in the family tree will have family ID data and parents ID. This allows machine learning technique and Bayesian analysis technique to determine if family tree or parents have high association with the DMD disease. One of the main advantages of Bayesian technique is that the probability of a disease occurrence can be estimated.

### *D. Data Splitting*

The data set were first divided into the training and testing sets in the ratio of 0.8:0.2. A validation set was a 20% split from the training set. Validating using a single validation set and testing using a single test set is considered to be unfair[5]. Ten-fold cross validation is applied in order to obtain average performance of machine learning algorithms and Bayesian analysis.

### *E. Model Selection*

The predictive performance of the widely used machine learning methods was compared. These methods include Bayesian analysis[9], Random Forest, SVM, Gradient Boosting and Neural Networks. Each model learned from the training data set and the learned models were first validated with the validation set. Once the best model was determined using accuracy performance, the best model is tested with the test data set 10 times according to the principle of ten-fold cross validation.

### *F. Model Evaluation and Prediction*

A confusion matrix was obtained for each machine learning model. The average accuracy, precision, recall and F-measure were calculated for each model[11]. In our case, we would like to have high precision and high recall. In order to help confirm the performance of our method, sensitivity and specificity is also computed.

### G. Variable Importance

Once the highest performing model is selected, the model parameters can be used to determine the variable importance. In addition, the association between variables are also determined. The variable importance can help improve the understanding how a variable or a feature affect the severity of the disease.

All methods were implemented using scikit-learn and keras packages[12-13]. The machine used to run all the codes has Intel Core i7 gen 12 processor with 256 GB RAM and Nvidia Quadro RTX 6000 GPU.

## RESULTS

After the machine learning techniques and Bayesian analysis were performed, the average accuracy, precision and recall for each technique were calculated and shown in Table 1. According to Table 1, the Gradient Boosting technique seems to perform best based on the value of the F-measure and accuracy. The Gradient Boosting also gave us the probability estimates for each test data which allowed us to determine the feature importance of the data set. Figure 2 shows an example of the feature importance analysis for our data set. We can see that the CK level and DMD gene detection are very significant to the precision of the disease.

Table 1: performance of DMD risk assessment techniques

| Technique | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Random Forest | 0.89 | 0.89 | 0.78 | 0.83 |
| SVM | 0.82 | 0.68 | 0.71 | 0.69 |
| Gradient Boosting | 0.91 | 0.93 | 0.91 | 0.92 |
| Neural Network | 0.89 | 0.83 | 0.77 | 0.80 |
| Bayesian method | 0.89 | 0.96 | 0.81 | 0.88 |


Figure 2: Feature importance ranking for the prediction with gradient boosting

However, when we test our findings with the physicians, we found that the physicians prefer to use Bayesian technique and Gradient Boosting together instead of one single method. This is probably due to the fact that the physicians were very familiar with genetic analysis based on Bayesian and they can understand the results of Bayesian technique more compare to the results of gradient boosting technique.

## CONCLUSION

We have described how to perform risk assessment for a genetic disease such as Duchenne Muscular Dystrophy (DMD). Our approach in using binary classification techniques using machine learning rather than traditional Mendelian method seems to be a correct decision. Gradient Boosting method appears to outperform other techniques including traditional approach. However, we manipulate a large portion of data, especially the data of healthy people, in order to balance the classes of the data set. In addition, family tree information in the form of drawing which is important to the understanding of the disease by the physicians were translated into text-based data and used as features in the prediction. The accuracy of the gradient boosting technique is the best among all binary classification techniques that we use in and it even outperforms the traditional method such as Bayesian method. The result of this work could be applied to other genetic diseases in order to improve risk assessment which leads to the prevention of the disease inheritance to the next generation especially for an uncurable disease such as DMD.

ACKNOWLEDGMENT

This research project is supported by Mahidol University.

## *REFERENCES*


[1] Bernardini, C., "Duchenne Muscular Dystrophy: Methods and Protocols," Humana, New York, NY, 2018.

[2] Robert Brooker, Genetics: Analysis & Principles, McGraw Hill; 7th edition (January 9, 2020).

[3] Njage, P., Henri, C., Leekitcharoenphon, P., Mistou, M., Hendriksen, R., Hald, T., "Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data," in Risk Analysis, vol. 39, issue 6., Wiley, 2019, pp. 1397–1413.

[4] Porras, A., Rosenbaum, K., Tor-Diez, C., Summar, M., Linguraru, M. G., "Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study," Lancet Digit Health. 2021.

[5] Molinaro, A., Simon, R., & Pfeiffer, R., "Prediction error estimation: A comparison of resampling methods," Bioinformatics, 21(15), 2015, pp. 3301– 3307.

[6] Machado, G., Mendoza, M. R., & Corbellini, L. G., "What variables are important in predicting bovine viral diarrhea virus? A random forest approach," Veterinary Research, 46(1), 2015, pp. 1– 15.

[7] Ogutu, J. O., Piepho, H.-P., & Schulz-Streeck, T., "A comparison of random forests, boosting and support vector machines for genomic selection," BMC Proceedings, 5(Suppl. 3), S11, 2011.

[8] Kuhn, M., "Building predictive models in R using the caret package," Journal of Statistical Software, 28(5), 2008, pp. 1– 26.

[9] Sang Medicine, "An Introduction to Risk Analysis in Inherited X-Linked Recessive Disorders", Practcal-Haemostasis, last access, 1 October 2022, https://practical-haemostasis.com/Genetics/bayesian_risk_analysis.html

[10] Theodoridis, S., "Machine Learning: A Bayesian and Optimization Perspective", Academic Press, 2nd edition, 2020

[11] Hicks, S.A., Strmke, I., Thambawita, V. et al., "On evaluation metrics for medical applications of artificial intelligence," Scientific Report 12, 5979, 2022.

[12] scikit-learn: machine learning in Python, https://scikit-learn.org

[13] keras: the Python deep learning API, https://keras.io