MAJORITY VOTING APPROACH FOR THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES TO UNDERSTAND GENDER-RELATED SKELETAL MUSCLE AGING

Abdouladeem Dreder[1], Muhammad Atif Tahir[1], Huseyin Seker[1] and
Muhammad Naveed Anwar[2]

Bio-Health Informatics Research Group
[1]Department of Computer Science and Digital Technologies
[2]Mathematics and Information Sciences
Faculty of Engineering and Environment
The University of Northumbria at Newcastle upon Tyne, NE1 8ST,
United Kingdom.
{a.dreder,muhammad.tahir,huseyin.seker,naveed.anwar}@northumbria.ac.uk

ABSTRACT

Understanding gene function (GF) is still a significant challenge in system biology. Previously, several machine learning and computational techniques have been used to understand GF. However, these previous attempts have not produced a comprehensive interpretation of the relationship between genes and differences in both age and gender. Although there are several thousand of genes, very few differentially expressed genes play an active role in understanding the age and gender differences. The core aim of this study is to uncover new biomarkers that can contribute towards distinguishing between male and female according to the gene expression levels of skeletal muscle (SM) tissues. In our proposed multi-filter system (MFS), genes are first sorted using three different ranking techniques (t-test, Wilcoxon and ROC). Later, important genes are acquired using majority voting based on the principle that combining multiple models can improve the generalization of the system. Experiments were conducted on Micro Array gene expression dataset and results have indicated a significant increase in classification accuracy when compared with existing system.

KEYWORDS

Multi-Filter System, Filter Techniques, Micro Array Gene Expression, Skeletal muscle

Jan Zizka et al. (Eds) : CCSEIT, AIAP, DMDB, MoWiN, CoSIT, CRIS, SIGL, ICBB, CNSA-2016
pp. 237–244, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60619

1. INTRODUCTION

Sexual dimorphism of skeletal muscle can occur due to age [1] and many of these age-related changes in skeletal muscle appear to be influenced by gender [2], [17], [18]. For example, the muscle mass of men is larger than that of women, especially for type II fibers, while the type I muscle fibers proportion of oxidative is higher in women [3]. Welle et al. reported that the muscle mass of men is larger than that of women [1], [11], [12], due to the higher level of testosterone and the anabolic effect of testosterone is well known. However, previous studies have failed to identify which genes are responsible for anabolic effects. The molecular biases related to gender difference are still fuzzy [1]; 50% of the cell mass of the human body is muscle, so skeletal muscle is considered an important issue. There are several changes in skeletal muscle related to age that seem to be influenced by gender [4]. These changes in gene expression could be responsible for the decline in muscle function [5]. In relation to sex, despite the fact that there are a higher number of genes in expression related to gender difference, very few genes can help to interpret the gender difference issue [3]. For the profiles of men and women, there are few comparisons of broad gene expression that have been carried out [5].

Janssen et al [13] reported that the reduction of skeletal muscle (SM) mass related to age starts in the third decade. This decrease starts to appear in the lower body SM. To find differences between men and women, they used t-test, pearson correlation and multiple regression to determine the relationship between age and skeletal muscle. Dongmei et al [2] used basic statistical analysis to make a comparison between males and females in each set of age using gene expression profiles from skeletal muscle tissue. They identified important sex and age related gene functional groups using intensity-based Bayesian moderated t-test and logistic regression. This was the first study that offers global proof for the occurrence of extensive sex changes in the aging process of human skeletal muscle. Although the study showed interesting results, but they had used genes belonging to X and Y chromosomes, which can easily discriminate genders. Experiments were conducted using 3 groups namely older women versus old men, young women versus older women, and young men versus older men. But the main problem with their study is that important genes are identified using whole training data. This can lead to poor generalization because one of the fundamental goal of machine learning is to generalize beyond the samples in the training data.

The main aim of this paper is to extend the work reported by Dongmei et al [2] by identifying important genes with good generalization ability. In our proposed approach which is basically inspired from ensemble of feature ranking methods for data intensive application [16], genes are first sorted using three different ranking techniques (t-test, Wilcoxon and ROC). Later, important genes are acquired using majority voting based on the principle that combining multiple models can improve the generalization of the system. The scope of this paper is the selection of the most reliable genes and the evaluation of classification power of selected genes. Experiments were conducted on Micro Array gene expression dataset and results have indicated a significant increase in classification accuracy when compared with the genes obtained by the system in [2]. Our proposed technique is able to identify differentially-expressed genes for the following three case studies in relation to age and gender differences

- Young Women versus Old Women
- Young Men versus Old Men
- Old Men versus Old Women

This paper is organized as follows. Section II describes material and the proposed method followed by results and discussion in Section III. Section IV concludes the paper.

2. MATERIAL AND PROPOSED METHOD

A. Micro array gene expression data set

In this study, the dataset contains a microarray dataset of gene expression of skeletal muscle arm tissue. Dataset is publicly available in the Gene Expression Omnibus (GEO) database [2]. The subjects comprise 22 healthy males and females of various ages, in which 7 males & 7 females are young (20-29 years old), and 4 males & 4 females are old (61-81 years old). The whole Ribonucleic Acid (RNA) was extracted and gene expression profiling was implemented utilising Affymetrix human genome U133 Plus 2chip. As in [2], this data set is divided into three cases, first case involve 11 females (7 young and 4 old), second case consists of 11 males (7 young and 4 old) and the last case contains 8 samples (4 old men and 4 old women).

B. Genes subset selection using Feature ranking techniques

Bioinformatics data have extremely high dimensionality. The above dataset consists of around 55,000 genes with only 22 samples. This is considered a significant challenge to machine learning methods. This means that there are a large number of features than samples. To address this problem, it is important to select a small relevant features subset to reduce processing time and to avoid over fitting [6]. The possible solution is the feature ranking methods. In this study, three different filter methods are investigated and are shown below

- **T-test**: a statistical hypothesis where the statistic follows a Student distribution [9]. It is usually used to evaluate if the averages of two classes are not statistically similar by computing the variability and difference between two classes.

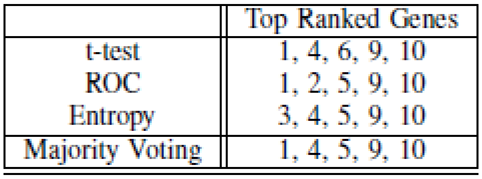
- **Entropy**: is normally used for high dimensional data to select the suitable number of features using the principle of Entropy.

- **ROC**: offers an active method to characterize the classifier sensitivity versus specificity.

C. Classification

Selected subset of genes are tested for its generalisation power using supervised classification. k-nearest neighbor (kNN) classifier (k=1,3) is used to used to evaluate the system performance. The leave-one-out cross validation (LOOCV) technique is used for evaluation.

Table 1 : Majority Voting to Select Important Genes.



	Top Ranked Genes
t-test	1, 4, 6, 9, 10
ROC	1, 2, 5, 9, 10
Entropy	3, 4, 5, 9, 10
Majority Voting	1, 4, 5, 9, 10

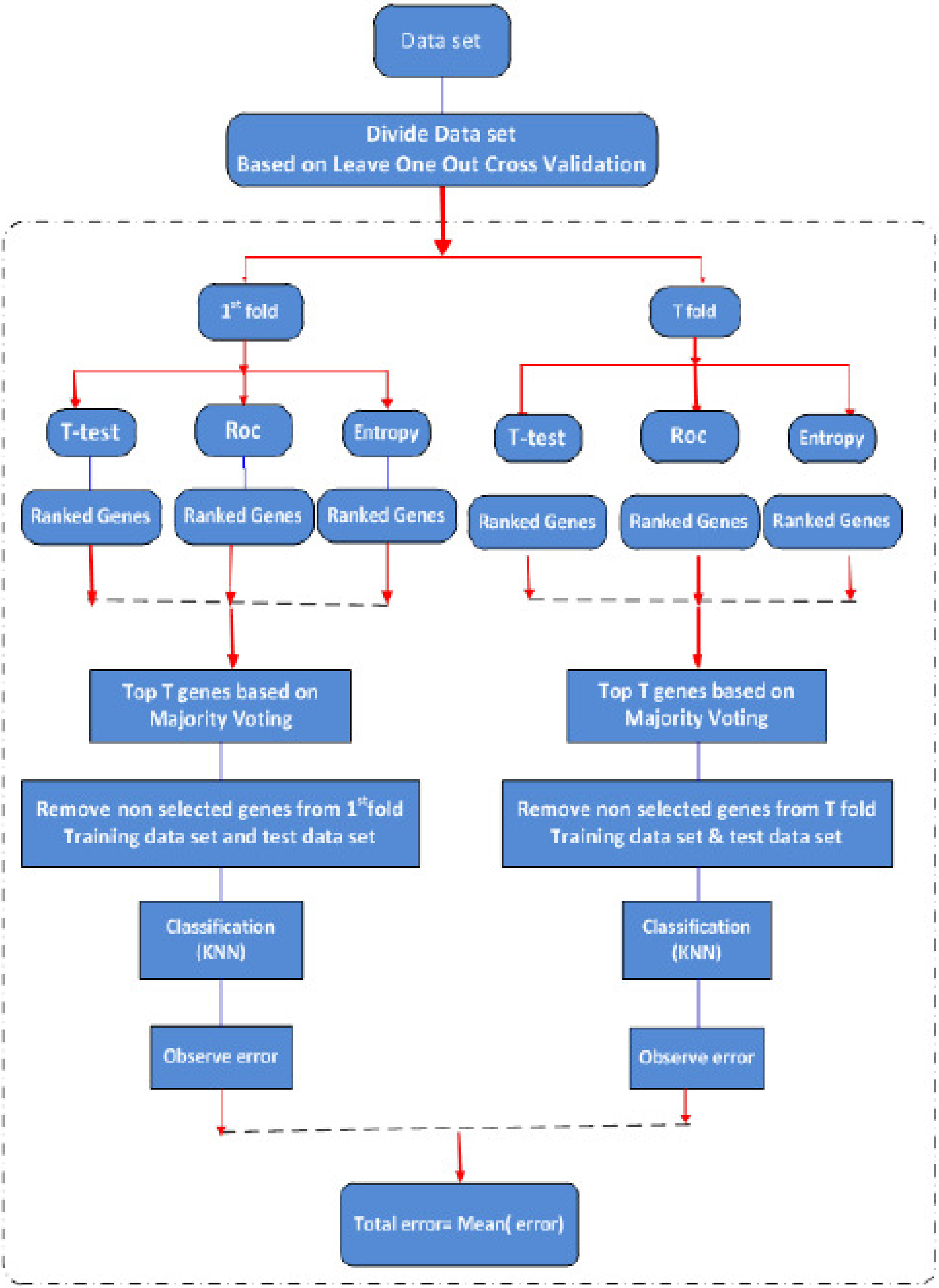

Data set
Divide Data set
Based on Leave One Out Cross Validation
1st fold
T fold
T-test
Roc
Entropy
Ranked Genes
Top T genes based on Majority Voting
Remove non selected genes from 1st fold Training data set and test data set
Remove non selected genes from T fold Training data set & test data set
Classification (KNN)
Observe error
Total error= Mean(error)

Fig. 1. Proposed Multi-Filter System (MFS).

1) *k-nearest neighbor*: The main objective of *k*-nearest neighbor (*k*-NN) classifier is to discover set of *k* objects in the training set that are similar to the objects in the test group [14]

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_{x_i} - a_{x_j})^2} \quad (1)$$

where *a* is the feature vector of x^{th} sample.

D. Proposed System

Figure 1 shows the framework of the proposed system which is inspired from the fact that combining multiple models can improve the generalization of the system. We first divided the data set using leave-one-out-cross validation into T folds. In other words, there are 20 folds for 20 samples where each fold consists of 19 samples for training and one sample for testing. For each fold multi filter system (MFS) is applied, which includes three different rank feature filters T-test, ROC and Entropy. Each filter ranking technique is responsible to sort genes according to criteria specified in the filter ranking methods. From these sorted genes, N unique subset of genes are obtained based on majority voting. This is clearly depicted using Table I. Lets assume that there are total 10 genes and the objective is to select top 5 genes. Genes 9, and 10 are selected by all feature ranking techniques so these are most important genes. Genes 1, 4, 5 are selected twice and thus are also considered as important genes by the system. It should be noted that due to majority voting, genes 2, 3 and 6 are not selected by the system. Later, kNN is applied on the new subset of genes in order to check the predictive performance.

3. RESULTS AND DISCUSSION

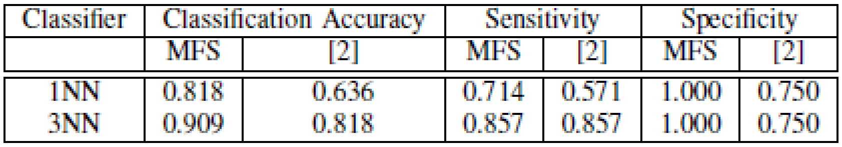
In this section, we will evaluate the performance of the multi-filter system (MFS). The proposed system is also compared with the system presented in [2], in which 75 genes are identified for three categories (male young versus male old, female young versus female old and male old versus female old) from total of 54623 genes. In order to have a fair comparison, the same number of genes are selected from MFS and compared with the genes identified in [2]. The evaluation metrics used in this study are: Classification accuracy, Sensitivity and Specificity.

A. Case Study 1: Young Men versus Old Men

This case study consists of 11 male samples (7 young and 4 old). Table II shows the performance of MFS when compared with the genes identified by Liu et al [2]. It is observed that the best performance is obtained using 3NN classifier which is 90.9% while genes obtained by [2] only able to achieve 81.8%. This improvement is mainly due to high specificity. Further analysis has revealed that out of 75 genes, only 9 genes are common in both systems. Some new genes are identified, that can play an important role in age differences of young and old males. Some of the new genes are shown in Table III along with 9 genes that are selected by both systems. These new genes can be very useful for biologist in order to identify the differences between young and old males.

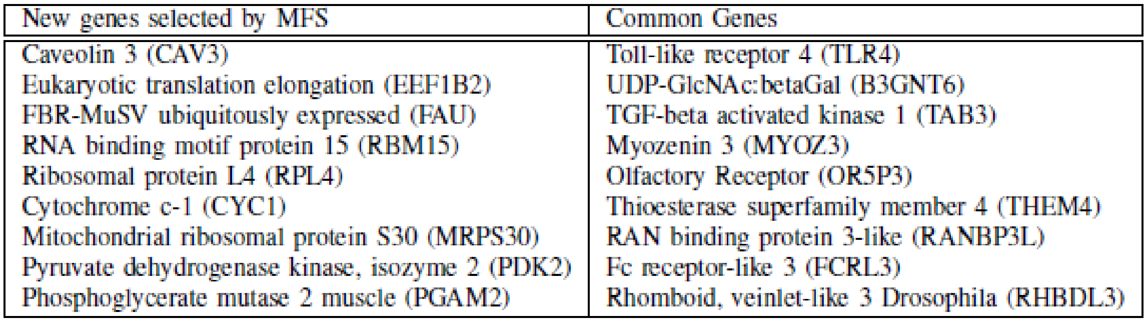
Figure 2 shows the performance of the system by varying the number of genes. It is observed that the best performance is obtained by using 10 or 20 genes and afterwards, there is a 10% drop in performance. This may be due to selection of some genes that can degrade the performance of the system. Future work aims to investigate wrapper techniques to identify these genes.

Table 2. Young Men Versus Old Men



Classifier	Classification Accuracy		Sensitivity		Specificity	
	MFS	[2]	MFS	[2]	MFS	[2]
1NN	0.818	0.636	0.714	0.571	1.000	0.750
3NN	0.909	0.818	0.857	0.857	1.000	0.750

Table 3. Young Men Versus Old Men. New Genes Selected by the Proposed System. Common Genes Selected by Proposed System and System by [2].



New genes selected by MFS	Common Genes
Caveolin 3 (CAV3)	Toll-like receptor 4 (TLR4)
Eukaryotic translation elongation (EEF1B2)	UDP-GlcNAc:betaGal (B3GNT6)
FBR-MuSV ubiquitously expressed (FAU)	TGF-beta activated kinase 1 (TAB3)
RNA binding motif protein 15 (RBM15)	Myozenin 3 (MYOZ3)
Ribosomal protein L4 (RPL4)	Olfactory Receptor (OR5P3)
Cytochrome c-1 (CYC1)	Thioesterase superfamily member 4 (THEM4)
Mitochondrial ribosomal protein S30 (MRPS30)	RAN binding protein 3-like (RANBP3L)
Pyruvate dehydrogenase kinase, isozyme 2 (PDK2)	Fc receptor-like 3 (FCRL3)
Phosphoglycerate mutase 2 muscle (PGAM2)	Rhomboid, veinlet-like 3 Drosophila (RHBDL3)

B. Case Study 2: Old Men versus Old Women

This case study consists of 8 adults (4 old men versus 4 old women). Table IV shows the performance of MFS when compared with the genes identified by Liu et al [2]. It is observed that genes selected using MFS have classification accuracy of 100% using both 1NN and 3NN with high Sensitivity and Specificity.

C. Case Study 3: Young Female versus Old Female

This case study consists of 11 female samples (7 young and 4 old). Table V shows the performance of MFS when compared with the genes identified by Liu et al [2]. Again, the best performance is obtained using 1NN classifier which is 91%. While genes identified by [2] are only able to achieve 72.2% which indicates the important improved generalisation ability of the proposed system. We argue that improvement in performance is mainly due to high Specificity as Sensitivity which is same in the both systems.

4. CONCLUSION

In this study, muti-filter system (MFS) is proposed to identify important genes for Males and Females using skeletal muscle. Genes are first sorted using three different ranking techniques (t-test, Wilcoxon and ROC). The proposed system is evaluated on publicly available microarray dataset of gene expression of skeletal muscle arm tissue. Later, important genes are acquired using majority voting based on the principle that combining multiple models can improve the generalization of the system. The results have indicated that the classification performance achieved by the proposed system yields the best classification performance when compared with similar number of genes identified in previous study [2]. Future work aims to improve the performance by identifying more important genes through Wrapper Feature Ranking techniques rather than filter based feature ranking techniques.

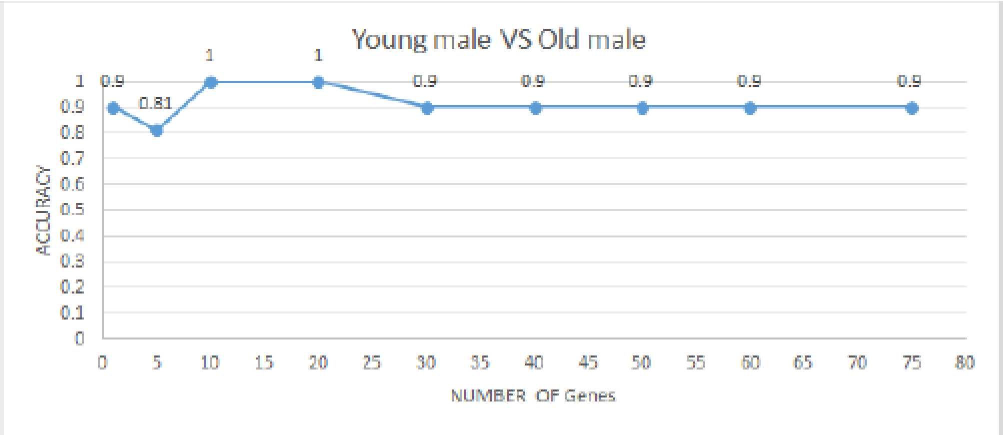


Fig. 2. Graph showing performance of the system using varying number of genes.

Table 4. Old Men Versus Old Women

Classifier	Classification Accuracy		Sensitivity		Specificity	
	MFS	[2]	MFS	[2]	MFS	[2]
1NN	**1.000**	0.750	1.000	0.500	1.000	1.000
3NN	**1.000**	0.375	1.000	0.500	1.000	1.000

Table 5. Young Female Versus Old Female

Classifier	Classification Accuracy		Sensitivity		Specificity	
	MFS	[2]	MFS	[2]	MFS	[2]
1NN	**0.909**	0.454	0.857	0.571	**1.000**	0.025
3NN	0.722	0.722	0.857	0.857	0.500	0.500

REFERENCES

[1] S. Welle, R. Tawil, and C. A. Thornton, “Sex-related differences in gene expression in human skeletal muscle”, PLoS One, vol. 3, no. 1, pp. e1385-e1385, 2008

[2] D. Liu, M. A. Sartor, G. A. Nader, E. E. Pistilli, L. Tanton, C. Lilly, et al., “Microarray analysis reveals novel features of the muscle aging process in men and women”, Biological Sciences, vol. 68(9), pp. 1035–1044, 2013

[3] D. D. Liu, M. A. Sartor, G. A. Nader, L. Gutmann, M. K. Treutelaar, E. E. Pistilli, H. B. IglayReger, C. F. Burant, E. P. Hoffman, and P. M. Gordon, “Skeletal muscle gene expression in response to resistance exercise: sex specific regulation”, BMC Genomics, vol. 11, no. 1, pp. 659, 2010.

[4] G. Sifakis, I. Valavanis, O. Papadodima, and A. A. Chatziioannou, “Identifying Gender Independent Biomarkers Responsible for human Muscle Aging Using Microarray Data”, Bioinformatics and Bioengineering (BIBE), pp. 1-5, 2013

[5] S. M. Roth, R. E. Ferrell, D. G. Peters, E. J. Metter, B. F. Hurley, and M. A. Rogers, “Influence of age, sex, and strength training on human muscle gene expression determined by microarray”, Physiological genomics, vol. 10, pp. 181-190, 2002.

[6] Y. Saeys, I. a. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics”, Bioinformatics, vol. 23, pp. 2507-2517, 2007.

[7] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, “RankGene: identification of diagnostic genes based on expression data”, Bioinformatics, vol. 19, pp. 1578-1579, 2003

[8] K. Murphy. “Machine learning: a probabilistic perspective”. Cambridge MA: MIT Press, 2012.

[9] N. Thouleimat, D. Hernandez-Lobato, and P. Dupont, “Variance Estimators for t-Test Ranking Influence the Stability and Predictive Performance of Microarray Gene Signatures”, European Conference on Computational Biology, 2010.

[10] S. Sahan, K. Polata, H. Kodazb, and S. Gne, “Anewhybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis”, Computers in Biology and Medicine, vol. 37, pp. 415-423, 2007.

[11] M. Visser, M. Pahor, F. Tylavsky, S. B. Kritchevsky, J. A. Cauley, A. B. Newman, B. A. Blunt, and T. B. Harris, “One-and two-year change in body composition as measured by DXA in a population-based cohort of older men and women”, Journal of applied physiology, vol. 94, pp. 2368-2374, 2003.

[12] V. A. Hughes, W. R. Frontera, R. Roubenoff, W. J. Evans, and M. A. F. Singh, “Longitudinal changes in body composition in older men and women: role of body weight change and physical activity”, The American journal of clinical nutrition, pp. 473-481, 2002

[13] I. Janssen, S. B, Heymsfield, Z. Wang, and R. Ross, “Skeletal muscle mass and distribution in 468 men and women aged 1888 yr”, Journal of applied physiology, vol. 89.1 pp. 81-88, 2000.

[14] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, “Top 10 algorithms in data mining”, Knowledge and information systems, vol. 14, pp. 1-37, 2008.

[15] A. C. Haury, P. Gestraud, and J.-P. Vert, “The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures”, PloS one, vol. 6, p. e28210, 2011.

[16] W. Altidor, T. M. Khoshgoftaar and J V Hulse and A. Napolitano, “Ensemble Feature Ranking Methods for Data Intensive Computing Applications”, Handbook of Data Intensive Computing, pp 349-376, 2011

[17] A. Y. Guo, K. S. LeunG, P. M. F. Siu, J. H. Qin, S. K. H. Chow, L. Qin, C. Y. Li, and W. H. Cheung, “Muscle mass, structural and functional investigations of senescence-accelerated mouse P8 (SAMP8)”, Experimental Animals, vol. 64, p. 425, 2015.

[18] R. R. Kalyani, M. Corriere, and L. Ferrucci, “Age-related and disease-related muscle loss: the effect of diabetes, obesity, and other diseases”, The Lancet Diabetes & Endocrinology, vol. 2, pp. 819-829, 2014.