Robust and Explainable Stage Prediction in Duchenne Muscular Dystrophy

Promita Ghosh[1], **Ragav Krishna**[2], **Lakshmi B. Raman**[2], **Malay Bhattacharyya**[1, 3*]

[1]Centre for Artificial Intelligence and Machine Learning, Indian Statistical Institute, Kolkata
[2]Molecular Diagnostics Counseling Care and Research Centre, Coimbatore, Tamil Nadu, India
[3]Machine Intelligence Unit, Indian Statistical Institute, Kolkata
promitaghosh12a@gmail.com, thenalirajaguru@gmail.com, drlakshmi@mdcrcindia.org, malaybhattacharyya@isical.ac.in

Abstract

Duchenne muscular dystrophy (DMD) is one of the life-threatening rare genetic disease affecting millions of male minors across the globe. Given its progressive nature, we can demarcate the various stages of DMD through the loss of muscular movements, ambulation, respiratory difficulties, and cardiac dysfunction. In this work, we employ machine learning models for understanding the progression of DMD through the prediction of its stages. Our attempts to predict the stages of DMD on the data collected by Molecular Diagnostics, Counseling, Care and Research Center (MDCRC) from 223 visits of 90 subjects demonstrate more than 80% accuracy with the state-of-the-art methods. We further study the biological/physiological importance of features in characterizing the stages of DMD.

Introduction

Duchenne muscular dystrophy (DMD) is a lethal pediatric genetic disorder primarily affecting male minors. It is a congenital myopathy attributed to the mutation in X chromosome that prevents the production of the muscle isoform of dystrophin (Duan et al. 2021). As per global statistics, approximately 20,000 children are diagnosed with DMD each year [1]. DMD is progressive in nature and the child loses its ambulation and other muscular movements with age. The affected individuals are wheelchaired within the age of ten years and finally succumb to the disorder in their early twenties due to respiratory distress and cardiac issues. Mortality in DMD patients rapidly increases with age (Llamosas-Falcón et al. 2022). The median life expectancy of subjects with DMD is hardly 22 years (Broomfield et al. 2021). The only gold standard of drug intervention is steroid administration and appropriate physiotherapy to handle progression, associated deformities, contractures and other related complications. Despite not being one of the deadliest diseases, DMD has a significant importance as a neglected disease. The issue at hand is critical as it affects minors and is progressive in nature.

Molecular Diagnostics, Counseling, Care and Research Center (MDCRC), through its long-term initiatives, has already identified and confirmed that average DMD minors in the state of Tamilnadu is slightly higher than the global average of 1 in 3500 live male births (Emery 1991). No cure is reported for this disorder as of now and it is a global problem without a solution. Regular clinical intervention including rehabilitation measures and continued observation are the only options to delay the effects of this disease. As prediction approaches might pave the way for such intervention (as we hypothesize), thereby improving the quality of life of kids and their families, we can finally converge towards better personalized clinical care and a strong base towards a cure in the long run. We aim to develop a learning model for accurate and systematic identification of the most effective clinical intervention for DMD subjects. We believe that this might also act as a generic model for other muscular dystrophies in future.

No pattern of DMD progression, which can help parents to be timely prepared and medicos to clinically intervene, is employable as a generic model till date. The available prediction mechanisms are rule-based and manual. The awareness about the disorder or the clinical expertise to understand the progression is also lacking. Through prediction of the progression of the disorder, followed by customized physiotherapy exercises and other intervention strategies we should be able to predict progression. The major research questions to answer here are: (i) Can we make better planning about clinical intervention based on the knowledge about prior responses of the subjects? (ii) Can we predict disease progression with a learnt model? We hypothesize that a prediction model would help us better understand the influences of clinical intervention and its effects on the progression of disease. In terms of validation, MDCRC already owns a reasonable amount of field data to be tested on. However, we should have the scope to continue the project at least for 3 years to really observe the precision of long-term predictability in practice.

We primarily prepare a prediction model for understanding the progression of DMD. Note that a further granularity of the stages exists in the data collected by MDCRC. The data collected by MDCRC, over a period of more than a decade, is sufficient enough for a long-term understanding of the disease progression. The data being available in the

*To whom the correspondence should be made.

[1]https://www.pfizer.com/disease-and-conditions/duchenne-muscular-dystrophy

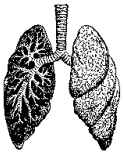
5-8 years	9-12 years	13-15 years	16-19 years	Above 20 years
				
Frequent falls; Calf muscle hypertrophy; Toe walking / Waddling gait; Gower's sign	Difficulty in getting up, Labored ambulation, Difficulty in stair climbing	No independent walk; Contracture of 1/2 joints; Able to manoevour wheelchair; Brooke's scale grade no more than 4	Contracture of no less than 3 joints; Weakness of upper limb and trunk muscles; Difficulty in maintaining erect posture and scoliosis; Respiratory difficulties	Increasing cardiac dysfunction; Heart failure; Death
Early Ambulant	**Late Ambulant**	**Early Non-ambulant**	**Late Non-ambulant**	

Table 1: Stages of DMD considered in the study. Note that pre-symptomatic was not considered as a separate stage due to the lack of appropriate data.

form of temporal clinical intervention of DMD subjects, the only option other than supervised learning appears to be a manual rule-based approach. However, this is highly dependent on the static data and prone to noise and biases.

The Research Problem

A prominent stage of DMD progression is not always definable for the subjects under study. The very first stage (of the progression of disease) is nothing but the diagnosis of the disease after the Pre-symptomatic stage. It is interesting to note that as the subjects do not start seeking medical intervention until the disease symptoms become visible, it is hard to collect data at the Pre-symptomatic stage. Hence, we can demarcate the major intermediate stages as Early Ambulant, Late Ambulant, Early Non-ambulant, and Late Non-ambulant (Birnkrant et al. 2018).

MDCRC owns data gathered through its grass-root level field work, its clinical observation sessions, and the related molecular data collected over a long time. There are many subjects for whom detailed clinical intervention data (along with demographic data and the molecular data) are already available and for cases, the period of clinical intervention is as long as 15 years. A portion of this data, which has much more detail related to clinical interventions, is available to enable the prediction of stages of DMD (Kumar et al. 2020). The prediction model might be helpful for precise prediction of DMD progression. Based on the predicted stage of progression of the disorder and the muscles affected, the model will provide an insight to the doctors and other medical staff to plan their intervention processes and therefore improve the quality of life.

Methods

We worked on the data collected from 90 subjects (male minors) who visited MDCRC 2-3 times for getting support over the past few years. The collected data comprises 129 features for the 223 visits of 90 DMD subjects. The data comprises many missing/unidentified values. We applied supervised learning models to classify the subjects into the stage of DMD they belong to based on 31 features, which were complete across all the subjects. Categorical features (in string format) in the dataset were converted into numeric (in integer format) values. Similarly, range values (e.g., Wrist Flexion, Shoulder Flexion, etc.) were converted into unique numeric values after ordering them by descending values. Thereafter, we employ traditional supervised learning models (Random Forest, Extra Tree, Gaussian Naive Bayes, Support Vector Machine, bagging) to predict the stages of DMD subjects. The associated parameters of the different supervised learning models were tuned through rigorous sensitivity analysis.

Model	Accuracy	Precision	Recall
Support Vector Machine	76.11 %	0.8696	0.5939
Gaussian Naive Bayes	76.12 %	0.7095	0.6960
Extra Tree	79.10 %	0.7564	0.7085
Bagging	79.10 %	0.7368	0.7210
Random Forest	83.58 %	0.7358	0.7371

Table 2: Performance of the different classification models in predicting the stages of DMD. The model parameters were chosen through rigorous sensitivity analysis.

Results

The performance of different classification models are highlighted in Table 2. We further employ the Random Forest classifier to calculate the importance of different features. The ranking of the top 25 features, where a lower rank reflects superiority, are highlighted in Table 3.

On carefully studying the top-ranked features and their biological/physiological significance, we found some characteristics to be unique to different stages of the disease progression. For example, wrist extension is the one of the upper limb movements associated with the Late Non-ambulant stage of DMD. We also observed that the number of inter-

Rank	Feature	Biological/Physiological Significance
1	Total	MDFRS score
2	North star assessment scales total score	Impact of disorder
3	Mobility	Measurement of movement
4	Basic ADL	Measurement of basic activities
5	Impairment	Measurement of additional activities
6	Ankle Dorsi Flexion (Lt)	Mobility of left ankle
7	Arm function	Arm strength
8	Ankle Dorsi Flexion (Rt)	Mobility of right ankle
9	Age at the time of MDCare	Starting of intervention
10	BROOKES SCALE Total Score	Functioning of upper extremities of body
11	Primary Index	First kid seen in the family
12	Knee Flexion (Lt)	Mobility of left knee
13	Knee Flexion (Rt)	Mobility of right knee
14	nth Visit	Progression of interventions
15	Hip Flexion (Lt)	Mobility of left hip joint
16	Elbow Flexion (Rt)	Mobility of right elbow
17	Hip Flexion (Rt)	Mobility of right hip joint
18	Elbow Flexion (Lt)	Mobility of left elbow
19	Total number of visits	Consistency of intervention
20	Total numbers of affected members in the family	Relate the genotype, phenotype with the familiar propensity
21	Wrist Extension (Lt)	Mobility of left wrist
22	Hip Abduction (Lt)	Mobility of left hip abduction
23	nth in the family	Positioning in family
24	Hip Abduction (Rt)	Mobility of right hip abduction
25	Wrist Extension (Rt)	Mobility of the right wrist

Table 3: Ranking of important features.

ventions has a key role to play in characterizing the stage of DMD. Moreover, mobility of different portions of the body (e.g., ankle, knee, hip joint, elbow and hip abduction) measured with angles (generally within a range of 0^o-180^o) are found to be good descriptors of the stages of the disease. Interestingly, many of the standard scales used for measuring the severity of muscular dystrophy emerge to be significant to distinguish the stags of DMD. This is reflected by the selection of North star assessment scales total score (employs a 17-item rating scale) and BROOKES SCALE Total Score (ranges between 1-6) within the top 10 features.

As a whole, these outcomes provide an in-depth and explainable insight toward developing an application that can leverage limited features to predict the stages of DMD. The selection of important features makes it feasible to deploy such an application in a resource-constrained scenario aiming a better global health.

Conclusion

The systematic clinical intervention we are proposing is expected to increase the social awareness about DMD. If we receive more responses from people in the next few years, we can be very sure that it has removed the social fear about DMD. Moreover, the average time taken for the subject to succumb to the disorder is known. This generally happens when they attain their twenties. If we observe a better performance among a set of target patients, we can be sure whether it has an impact on the society. As DMD is a progressive disease, it will take several years to verify a true impact. The Model would be getting input time series data which might help in additional prediction. It is interesting to note that, as this work is pertaining to a disease progression, explainability is a demanding concern of the model. In this work, we commit more on rehabilitation measures and to be honest it is solely the consent of the parents whether they wish to pursue the prediction cues provided by the model. In future, we plan to extend this work to study the effect of lifestyle parameters like dietary habits (Bhattacharyya 2015; Bhattacharyya, Maity, and Bandyopadhyay 2017) on the healthcare aspects of DMD. It is truly challenging to ensure health equity to the people across the globe when such latent effects are functional. Moreover, it is important to contribute to data, models, experiments, team science, and public trust with collective efforts to help policymakers devise effective plans to combat with future health burdens (Grieve et al. 2023).

Acknowledgments

The authors gratefully acknowledge the financial support received from Google as a part of their “AI for Social Good” program.

Authors’ Contributions

LBR and MB formulated the problem together and designed the experiments. RK (with his team) collected the data. PG and RK carried out the data processing, run the experiments, and analyzed the data. LBR and MB prepared the interpretation of results. Rest of the authors have critically reviewed

the results and provided important feedback. PG and MB drafted the first version of the manuscript. All the authors read and approved the final version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Availability

Data and source codes will be made available on reasonable request.

References

Bhattacharyya, M. 2015. Disease dietomics. *XRDS: Crossroads, The ACM Magazine for Students*, 21(4): 38–44.

Bhattacharyya, M.; Maity, S.; and Bandyopadhyay, S. 2017. Exploring the missing links between dietary habits and diseases. *IEEE Transactions on NanoBioscience*, 16(3): 226–238.

Birnkrant, D. J.; Bushby, K.; Bann, C. M.; Alman, B. A.; Apkon, S. D.; Blackwell, A.; Case, L. E.; Cripe, L.; Hadjiyannakis, S.; Olson, A. K.; et al. 2018. Diagnosis and management of Duchenne muscular dystrophy, part 2: respiratory, cardiac, bone health, and orthopaedic management. *The Lancet Neurology*, 17(4): 347–361.

Broomfield, J.; Hill, M.; Guglieri, M.; Crowther, M.; and Abrams, K. 2021. Life expectancy in Duchenne muscular dystrophy: reproduced individual patient data meta-analysis. *Neurology*, 97(23): e2304–e2314.

Duan, D.; Goemans, N.; Takeda, S.; Mercuri, E.; and Aartsma-Rus, A. 2021. Duchenne muscular dystrophy. *Nature Reviews Disease Primers*, 7(1): 13.

Emery, A. E. 1991. Population frequencies of inherited neuromuscular diseases—a world survey. *Neuromuscular disorders*, 1(1): 19–29.

Grieve, R.; Yang, Y.; Abbott, S.; Babu, G. R.; Bhattacharyya, M.; Dean, N.; Evans, S.; Jewell, N.; Langan, S. M.; Lee, W.; et al. 2023. The importance of investing in data, models, experiments, team science, and public trust to help policymakers prepare for the next pandemic. *PLOS Global Public Health*, 3(11): e0002601.

Kumar, S. H.; Athimoolam, K.; Suraj, M.; Das Christu Das, M. S.; Muralidharan, A.; Jeyam, D.; Ashokan, J.; Karthikeyan, P.; Krishna, R.; Khanna-Gupta, A.; et al. 2020. Comprehensive genetic analysis of 961 unrelated Duchenne Muscular Dystrophy patients: Focus on diagnosis, prevention and therapeutic possibilities. *Plos one*, 15(6): e0232654.

Llamosas-Falcón, L.; Sánchez-Díaz, G.; Gallego, E.; Villaverde-Hueso, A.; Arias-Merino, G.; Posada De la Paz, M.; and Alonso-Ferreira, V. 2022. A population-based study of mortality due to muscular dystrophies across a 36-year period in Spain. *Scientific Reports*, 12(1): 3750.