Title : Advanced Machine Learning Techniques for Diagnosing Muscular Dystrophy: A Comprehensive Study of Feature Selection, Data Preprocessing, and Model Evaluation

Author(s): Ritvik Ellendula

Affiliation(s): There are no competing financial or personal interests and affiliations.

Funder(s): No funding was received from any outside party and consent for data collected was also received via the origin of the datasets.

Abstract :

Muscular dystrophies are life-defining ailments that dramatically affect muscle strength and result in drastic shifts to muscular condition, consequently leading to much shorter life spans for patients diagnosed with muscular dystrophy. However, some times, diagnoses could be made prior to the onset of physical symptoms for muscular dystrophy, which emphasizes the importance of creating methodologies of genetic diagnoses. For this methodology of diagnosis, machine learning seems optimal due to its ability to handle masses of data with extreme efficiency along with its accuracy and varying types of models. Specifically, models, such as Logistic Regression, Naive Bayes, K-Neighbors, and Random Forest seem to be very potent in diagnosing muscular dystrophy based on genetic information - gene expressions. By looking into the correlation and causation of different expressed genes, more insights into them will arrive, leading to better ways of diagnosing and treating muscular dystrophy.

Keywords : Muscular Dystrophy, Machine Learning, Diagnosis, Random Forest, Neuromuscular disorders

Introduction

Neuromuscular disorders (NMD) in their entirety is a term used to generally overview a heterogenous group of genetic conditions, which 1) have a specific type of genetic inheritance and 2) are encompassed by muscle degeneration and overall muscle weakness. Specifically, regarding the first point of genetic inheritance, these neuromuscular disorders often follow the inheritance pattern of autosomal dominant, recessive, or X-linked inheritance. Commonly, these disorders directly impact skeletal muscle due to causing motor nerve, neuromuscular junction, and cellular matrix irregularities. These neuromuscular disorders impact peripheral nerves and skeletal muscles as well, proving their depth and the extent to which they can damage the body as a whole. With this extensive nature, there are several different established “categories” to neuromuscular disorders as a whole, with extensive research being done distinguishing each

category from one another. Consequently, while all of them differ in some regards, especially in terms of genetics and specific protein translation, all of them are associated with clinical myopathy, leading to muscular weakness due to muscle fiber dysfunction.

Motor Neurone Disease	Amyotrophic Lateral Sclerosis, Spinal Muscular Atrophy, Spinal Bulbar Muscular Atrophy
Peripheral Neuropathies	Charcot-Marie Tooth Disease, Friedreich's Ataxia, Dejerine-Sottas Disease
Disorders of Neuromuscular Transmission	Myasthenia Gravis, Eton-Lambert Syndrome, Congenital Myasthenic Syndrome
Muscular Dystrophies	Duchenne, Becker, Limb-Girdle, Fascioscapulohumeral, Emery-Dreifuss, Oculopharyngeal, Distal, Congenital and Myotonic Dystrophy
Metabolic and Mitochondrial Myopathies	Lactate Dehydrogenase Deficiency, Carnitine Deficiency, Mitochondrial Myopathy, Phosphofructokinase Deficiency, Acid Maltase Deficiency, Phosphorylase Deficiency, Debrancher Enzyme Deficiency
Non-Dystrophic Myotonias	Myotonia Congenita, Paramyotonia Congenita, Periodic paralysis, Central Core Disease, Myotonia fluctuans

Figure 1. Types of Neuromuscular Disorders

Specifically, one of the most devastating types of neuromuscular disorders is muscular dystrophy and its many subcategories. Akin to most neuromuscular disorders, muscular dystrophy causes general muscle deterioration and weakness, but in particular, this inherited disease also contributes to the wasting of muscle tissue with each type being associated with different levels of strength loss and disability rate. Furthermore, each different type of muscular dystrophy also differs in terms of muscles impacted, age of genesis, and rate of prognosis.

Becker	Teen to early adulthood	Symptoms are almost the same as Duchenne, but less severe. It progresses more slowly than Duchenne. Survival goes into middle age. Becker disease is almost always limited to males. This is the same as with Duchenne.	**Duchenne**	Ages 2 to 6	Symptoms include general muscle weakness and wasting. It affects the pelvis, upper arms, and upper legs. Over time, it includes all voluntary muscles. Survival beyond the 20s is rare. It happens mostly in boys. Very rarely it can affect women, who have much milder symptoms and a better prognosis.

Table 1 Most common forms of muscular dystrophy (Becker's and Duchenne's)

However, a main commonality with all types of muscular dystrophies are their exclusive natures to males, due to the nature of the ailment. All types of muscular dystrophy are hereditary, similar to some neuromuscular disorders. While there are some cases where mutations occur to an otherwise healthy individual, causing an onset of muscular dystrophy, this is a more rare scenario. Therefore, regarding this hereditary aspect of muscular dystrophy specifically, it's primarily caused due to males inheriting genes in a different method than females. For instance, any child inherits two copies for every single gene - one from each parent. Therefore, if a parent has a mutated gene, that can be passed onto the offspring. More specifically though, muscular dystrophy can be inherited via recessive inherited disorder, dominant inherited disorder, or sex-linked X disorder. Sex-linked X disorder inheritance patterns are primarily the reason that males are primarily affected by types of muscular dystrophy.

As mentioned before, when inheriting genes from parents, the mechanisms for males and females differ vastly. A male obtains one X chromosome and one Y chromosome, while a female solely has two X chromosomes. However, that aspect of possessing only one X chromosome, makes males more susceptible to risk of inheritance or genesis of muscular dystrophy. For instance, if one of the genes on the X chromosome is mutated with muscular dystrophy, it means that the male will be impaired by the ailment. However, for a female, if one of the X chromosomes becomes mutated, they still have an extra chromosome, which can "mask" the effects or the initial mutation.

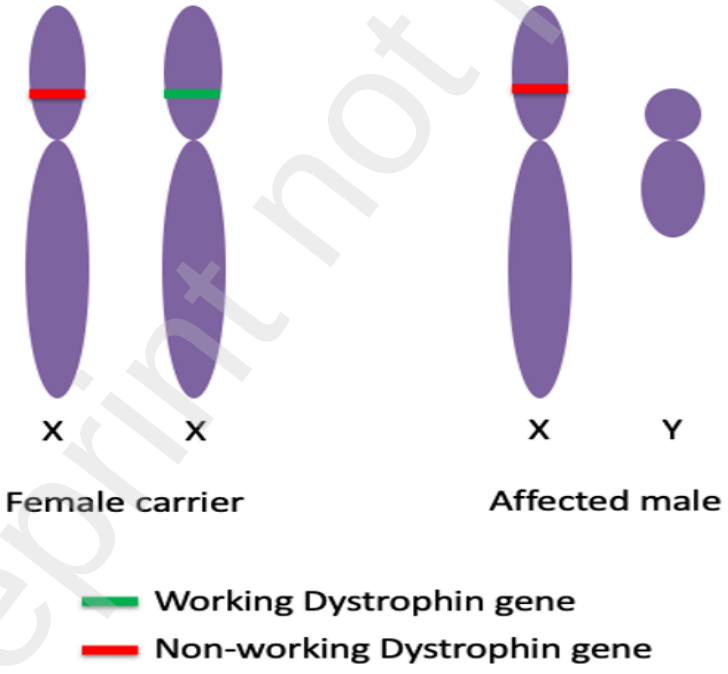

Figure 2 : Example of Dystrophin Gene impact on males vs females

The specific etiology of muscular dystrophy, apart from having an increased inheritance rate for males rather than females, appears in a specific X-linked gene: DMD. DMD is a specific gene that encodes for the production of dystrophin, and a mutated DMD gene marks the most prevalent cause for the onset of muscular dystrophy. Furthermore, this DMD gene is the largest in the human genome with 79 separate exons, meaning that this gene is the most prone to being mutated due to random spontaneous mutations. This spontaneity of the gene along with the fact that insufficient dystrophin levels may also cause muscle fibers to decay, contributes to the alarming number of 250,000 people in the United States being impaired by muscular dystrophy.

In recognition of the devastating impacts of muscular dystrophy, specifically some of the more prevalent like Duchenne's and Becker's, more effective methods of diagnosis along with better understanding of genes that might be responsible for the development of muscular dystrophy is essential to preventing this neuromuscular disease from spreading its grasp to more people than it already has. Specifically, machine learning algorithms paired with bioinformatic computing seem to offer a realm of possibility towards fighting the proliferation of muscular dystrophy through developing metrics of effective pre-emptive diagnosis.

Materials and Methods

Dataset Selection

For any machine learning algorithm, an appropriate amount of data is essential, specifically a strong amount of features to correspond with possible labels. Therefore, determining proper datasets for developing a diagnosis machine learning algorithm would be critical. Specifically, these datasets should relate heavily to muscular dystrophy with their being subjects with varying types of muscular dystrophy, such as Duchenne, Becker, and more along with "controls" or healthy subjects who have not been diagnosed with muscular dystrophy.

After an extensive examination of 50+ datasets on GEO (Gene Expression Omnibus), the optimal datasets were narrowed down to two choices - GSE3307 and GSE6011. These datasets fulfill the established criteria, along utilizing an experiment type defined as, "expression profiling by array". This description entails a technique used to measure the activity levels of thousands of genes at once using a DNA microarray, which helps researchers understand which genes are active in different conditions.

By selecting these datasets, we ensure that the machine learning algorithm has access to a wide range of genetic data, facilitating the identification of patterns and markers specific to different types of muscular dystrophy. Furthermore, the use of expression profiling by array in both datasets guarantees that the data is comparable, standardized, and suitable for integration into a unified model.

Overall, the chosen datasets provide a solid foundation for developing a machine learning algorithm capable of accurately diagnosing various types of muscular dystrophy, ultimately contributing to improved patient outcomes and a deeper understanding of the disease.

Context of Datasets

The first dataset, GSE 3307, uses comparative profiling for 13 different muscle groups. According to the dataset itself, " Groups studied are: Normal human skeletal muscle, Acute quadriplegic myopathy (AQM; critical care myopathy), Juvenile dermatomyositis (JDM), Amyotophic lateral sclerosis (ALS), spastic paraplegia (SPG4; spastin), Fascioscapulohumeral muscular dystrophy (FSHD), Emery Dreifuss muscular dystrophy (both X linked recessive emerin form; autosomal dominant Lamin A/C form), Becker muscular dystrophy (partial loss of dystrophin), Duchenne muscular dystrophy (complete loss of dystrophin), Calpain 3 (LGMD2A), dysferlin (LGMD2B), FKRP (glycosylation defect; homozygous for a missense mutation). "

The main goal of this dataset listed by the researchers who created it was the intention of discovering commonalities between separate diseases and their biochemical pathways. For instance, if Duchenne muscular dystrophy and Calpain 3 shared any genes that could potentially contribute to the development of one of the diseases.

Meanwhile, the second dataset, GSE6011, uses expression data primarily from the quadricep muscle of patients with DMD and age matched controls. Specifically, this dataset makes an important consideration, mentioning, "increased serum CK level and abnormal muscle histology are always present, boys with DMD are phenotypically indistinguishable from the normal ones at birth and, in their first years of life, acquire early motor milestones at normal times. A clear defect in muscle function becomes generally apparent by the end of the second year. As the disease is typically diagnosed between the ages of 3 and 7, the first two years are often considered and referred to as clinically presymptomatic." This also contributes to the necessity of a strong form of genotypically diagnosing DMD, as it would allow for a more precise methodology than doing so phenotypically.

The methodology of gathering the data samples in specific was through muscle biopsies, which means the removal of small muscle tissues, in this scenario from the quadriceps, to examine for gene expression data.

As a whole, these datasets both work very well in tandem, due to the sheer amount of data, which poses ample for machine learning algorithms. However, it's important to examine the specific contents and "shape" of each dataset. This is important for understanding future steps to go in the actual programming process, especially with pre-processing the data and merging the datasets together. By applying a simple shape function for each gene expression data, imported from GEO, we can see the disparity in the rows and columns present in each

dataset. This significant difference will need to be remedied in order to create one "whole" dataset.

```
Dataset Number Rows,Columns
GSE 3307 has: (44760, 242)
GSE 6011 has: (22283, 37)
```

Figure 3: Rows and Column Numbers for each dataset (Rows represent total number of genes looked at, while columns are number of subjects.

Description of Used Algorithms

The main language used throughout this entire machine learning process was Python, being conducted in Google Colab, with the purpose of permitting for more-open access to this research and the content discovered

Pre-Processing:

1. Created function for preprocessing (which allows multiple datasets to be pre-processed at once)
2. Applied log function on the expression matrix
3. Applied imputer (replacing extreme values with medians)
4. Ran imputer through the function, allowing the datasets to be normalized

Ordinary Least Squares Regression (OLS Regression):

1. OLS Regression relies on having an alternative "goal" variable, also known as a label, which in this case was our condition variable.
2. The OLS Regression works by trying to minimize the sum of squared differences between an observed value and an actual value, which in this case would mean minimizing the space between the label predictions and actual outcomes, often derived by the formula :
   $Y_i = B_0 + B_1 X_1 + E_i$
   Y(Dependent Variable); B(Linear Component); E(Error Value)

General Process

Dataset

First, the major essential step was loading each dataset from GEO. In order to do this, multiple modules were needed to be installed, such as GEOparse, pandas, etc. After this, the metadata and shape of each dataset were examined to gain a greater understanding of the data

being worked with. Gaining a fundamental knowledge of the datasets helped to ensure that it would be generally feasible to compare and merge both datasets.

```
!pip install GEOparse
!pip install combat
!pip install scipy
import GEOparse
#downloading the module that will allow us to load data from Genetic Omnibus
import pandas as pd
#downloading pandas for making arrays and accessing data
import numpy as np
#pretty similar to pandas
from sklearn.preprocessing import StandardScaler
#used in pre-processing for helping standardize data with multiple datasets
from sklearn.impute import SimpleImputer
#used in pre-processing for replacing missing values
from combat.pycombat import pycombat
#batch effects

#Feature selection packages
from statsmodels.stats.multitest import multipletests
import statsmodels.api as sm
from statsmodels.formula.api import ols
from sklearn.feature_selection import SelectKBest, chi2
import pandas as pd
from scipy.stats import ttest_ind

#Machine Learning Algorithms
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
```

Figure 4. List of all modules imported

Data Extraction & Filtering

After this initial examination, the data extraction process, which would take and create a pandas DataFrame from imported datasets, began. An expression matrix was made, which is a table that represents the levels of gene expression across different samples. It is a fundamental data structure in genomics and is used for various analyses like identifying differentially expressed genes, clustering samples, and building predictive models.

Doing this, added another layer of preparation for future steps, allowing for easier manipulation of the data and further benefit for data analysis. Additionally, another check was done to ensure that the data was not heavily modified by anything in the process, verifying the number of samples in each GSE dataset. However, after extracting the data and creating panda DataFrames, the head function was used to try to visualize how our data looks in each respective dataset for the moment. Immediately, a clear discrepancy was noticeable - the NaN values that some of the samples possessed. Therefore, this would need to be remedied at a future stage to ensure dataset functionality.

	GSM74244	GSM74245	GSM74246	GSM74247	GSM74248	\
ID_REF						
1007_s_at	1986.621216	2423.252197	3030.981689	2691.871582	2191.285400	
1053_at	308.545441	205.691345	400.727661	186.328323	309.432556	
117_at	305.663177	277.269165	322.180786	300.988678	351.027740	
121_at	3346.971680	2559.171875	3803.663330	3449.002930	3723.652832	
1255_g_at	116.961731	155.374146	127.237595	98.381233	186.740036	

	...	GSM121406	GSM121407	GSM121408	GSM121409	GSM121410	\
ID_REF	...						
1007_s_at	...	NaN	2254.563965	NaN	NaN	NaN	
1053_at	...	NaN	159.442871	NaN	NaN	NaN	
117_at	...	NaN	435.373108	NaN	NaN	NaN	
121_at	...	NaN	2624.151611	NaN	NaN	NaN	
1255_g_at	...	NaN	131.276306	NaN	NaN	NaN	

Figure 5. Discrepancy of NaN data values

However, prior to that modification, another step included filtering each dataset to ensure that they were examining the same genes specifically. For instance, GSE3307 had significantly more data, and along with that, more types of genes were analyzed than GSE6011. This means that completely different genes were included, which would make it difficult to merge each dataset together. Therefore, after the overlap percentage in specific genes was established, both were filtered to only contain genes that matched with each other, leading to a different amount of rows present in each dataset comparatively to the beginning of the coding process.

```
Shape of filtered GSE3307: (22283, 242)
Shape of filtered GSE6011: (22283, 37)
```

Figure 6. New Shape of Filtered Data

Pre-Processing

Finally, pre-processing of the data began. Preprocessing data is critical for strong data analysis, as it adjusts for differences in sequencing depth, sample preparation, and other technical factors that can affect measured expression levels. For instance, if in GSE3307 gene expressions were measured from a scale of 1-10 and in GSE6011 gene expressions were measured on a scale of 1-1000, it would cause extreme chaos and confusion when actually analyzing the data itself. Therefore, before that even begins, in the pre-processing, it is essential to standardize the data effectively for future stages. The process of normalization sought to bring the datasets to a common scale for a more fair comparison.

For this preprocessing process, a logarithm function was used in order to help with data standardization. Next, by finding the IQR (Interquartile Range), a methodology to remove outliers was established. This allowed the usage of the formula of Q1-1.5*IQR and Q3+1.5*IQR, which helped filter the expression matrix. Afterwards, an imputer was utilized in order to replace outliers with median and mean values.

After the standardization and preprocessing was finalized, a sequence to combine the two initial expression matrices was conducted, which allowed the creation of one large dataset to be used for the passing of several algorithms and programs.

```
Number of rows in combined_data: 21908
Number of columns in combined_data: 279
```

	GSM74239	GSM74240	GSM74241	GSM74242
0	10.612951	10.809627	11.007960	11.097372
1	5.774662	8.641031	8.900495	8.161290
2	8.919957	8.320556	7.870784	8.564891
3	11.572460	11.819829	11.460138	11.437432
4	7.704983	7.262651	6.653807	7.067918

Figure 7: Combined Data Matrix

Condition Assignment

The next important step was assigning conditions, which would act as labels for the features of the codes. Essentially, by adding conditions, such as “affected” and “normal” the program and machine learning algorithms especially would be able to tell if its utilization of the features would accurately assess the state of the given person. This step proved essential, especially due to the core goal: diagnosing or creating a methodology to diagnose muscular dystrophy as well as potentially finding possible genes related to the genesis of muscular dystrophy. Therefore, the feature process often involves using alternate variables, which in this case the dataset consisted of, and then having an outcome “goal” variable, also known as a label. In this case, the several variables in our dataset were our “features”, while the end goal of the diagnosis: affected or healthy was the label.

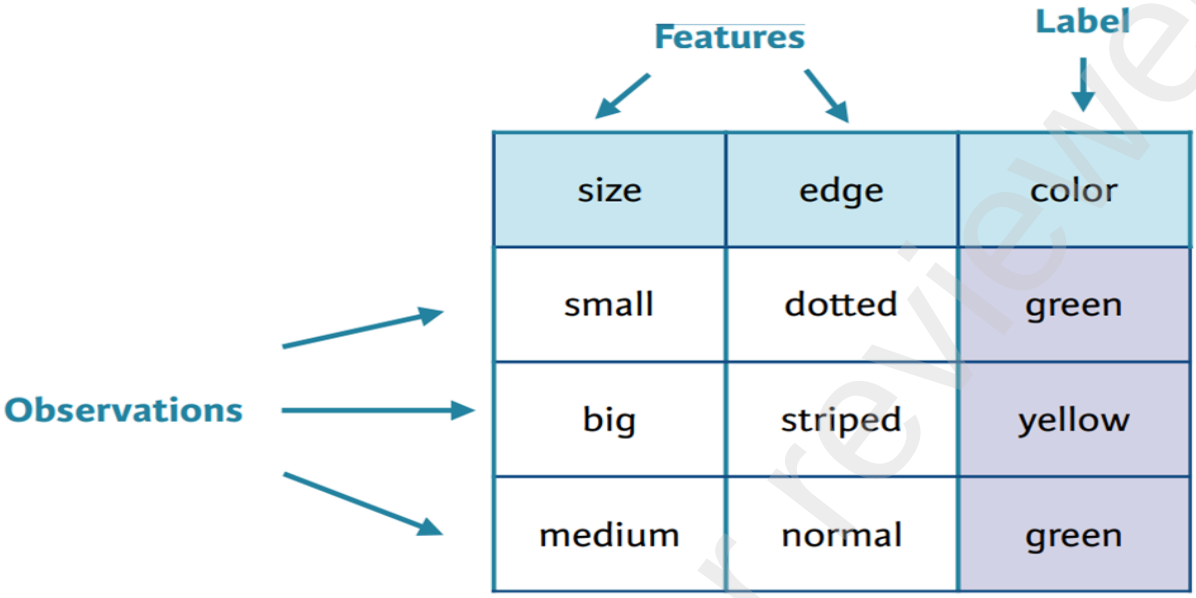

Table 2 : Examples of features and association with the label for a different scenario.

Gene Identification

Afterwards, an OLS regression model was run to gain an idea of genes of key interest that might want to be focused on more heavily for our future steps, especially in the machine learning process. For this OLS regression, a for loop was ran to iterate through the sequence of all possible genes in the experimental data (transposed version of the combined data matrix), which allowed fast checking of probability and role all of the genes played in contribution to the label variable(affected or healthy status).

```
Gene: 21290, p-value: 0.326, Coefficient: 2.256
Gene: 21291, p-value: 0.630, Coefficient: 0.902
Gene: 21292, p-value: 0.501, Coefficient: 1.644
Gene: 21293, p-value: 0.195, Coefficient: 3.741
Gene: 21294, p-value: 0.363, Coefficient: 0.957
Gene: 21295, p-value: 0.007, Coefficient: -2.023
Gene: 21296, p-value: 0.396, Coefficient: -2.297
Gene: 21297, p-value: 0.287, Coefficient: 2.124
Gene: 21298, p-value: 0.526, Coefficient: 3.119
Gene: 21299, p-value: 0.480, Coefficient: 1.639
Gene: 21300, p-value: 0.103, Coefficient: -2.035
Gene: 21301, p-value: 0.004, Coefficient: -6.944
```

Figure 8: Example output for genes “21290-21301” with p-values and coefficients

Derived from this plethora of outputs of genes,p-values, and coefficients, a list of the top k features, which in this case k=10, was created, allowing for some more specific features to be used for model training. This helps ensure that the model doesn’t get “overloaded” by the sheer amount of data, also making sure that we prioritize the parts of the dataset that seem to have the heaviest influence.

```
# Select top k features (e.g., k=10)
top_k_features = results_df.head(10)['gene'].tolist()

print(top_k_features)
```

```
[2259, 19471, 15894, 20481, 636, 34, 12206, 13227, 8088, 11086]
```

Figure 9. List of top 10 genes with the highest probability of impacting status (muscular dystrophy or healthy)

Algorithmic Testing

The final and most critical step for actually establishing and accomplishing the main goal of this project was to run the machine learning algorithms. The main algorithms tested were Logistic Regression, Naive Bayes, K Neighbors, and finally Random Forest. Prior to actually testing each model though, important steps included creating a test(20%) and training(80%) split for each model to be trained on with the input variable being the 10 “k” feature (genes) and the output variable being the label (condition). In this preparation for the machine learning stage, a definition function was established, “evaluate_model()”, which fitted the model and used the predictions also calculating other metrics, such as Accuracy, Precision, Recall, and F1 Score.

Specifically, the models utilized represent an integral part of the experiment, as different models perform better at different tasks for varying reasons. To better understand this divergence in performance it is crucial to gain a fundamental knowledge of the mechanisms behind different models and their methodologies.

First, logistic regression is a supervised machine learning model that fulfills binary classification tasks, which are the act of choosing between two different options through predicting probability of an occurrence. In logistic regression, there are two possible outcomes: yes/no, 0/1, or true/false. Specifically, logistic regression uses a certain function called a sigmoid function, which maps an arbitrary value on a plane ranging from 0 to 1. This sigmoid function is similar to the shape of an S, and converts any real value into this new output from 0 to 1. By nature, logistic regression works very well in situations where the dependent variable, which in our case is the status of the patient(healthy or affected) only has two outcomes. Furthermore,

through removal of large outliers, logistic regression is further strengthened, which was completed during pre-processing.

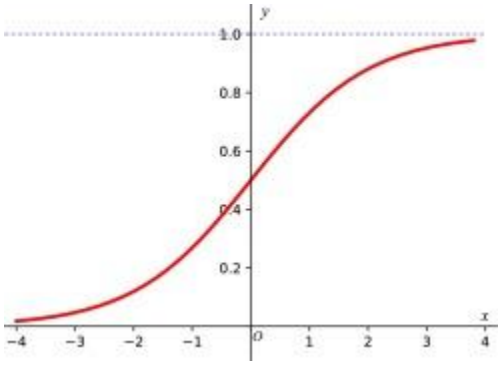

Figure 10. Sample Logistic Regression Curve

Second, Naive Bayes assumed that features are independent, which is why the algorithm is deemed as "Naive", using an assortment of Bayes' Theorem. In particular, Naive Bayes is used for classification problems, such as text classification, where data contains high dimensions, such as each extra character representing another additional input for the data. Furthermore, it's often used in spam filtering and rating classification.

Third, K-Neighbors works by using "proximity" to make classifications and predictions about grouping data. Theoretically, it can be used for both regression or classification, but typically it performs best at classification problems. K-Neighbors identifies the nearest neighbors of a point in space, assigning a "class" to that point. Essentially, it uses points nearby another point to determine and classify the original point. First, the "K" is identified, or the number of nearest neighbors that needs to be considered. After, the distance between possible neighbors is conducted. Then, the end involves finding these neighbors.

Lastly, the Random Forest model uses several decision trees, which have distinct weights and structures, to reach a single result. Essentially, decision trees constantly narrow possibilities down until there is one evident answer. Decision trees are often used in classification, but might be prone to bias and overfitting. Random Forest in particular generates a random subset of features, meaning that the random forest algorithm only uses a subset of features, while a decision tree algorithm would use all features.

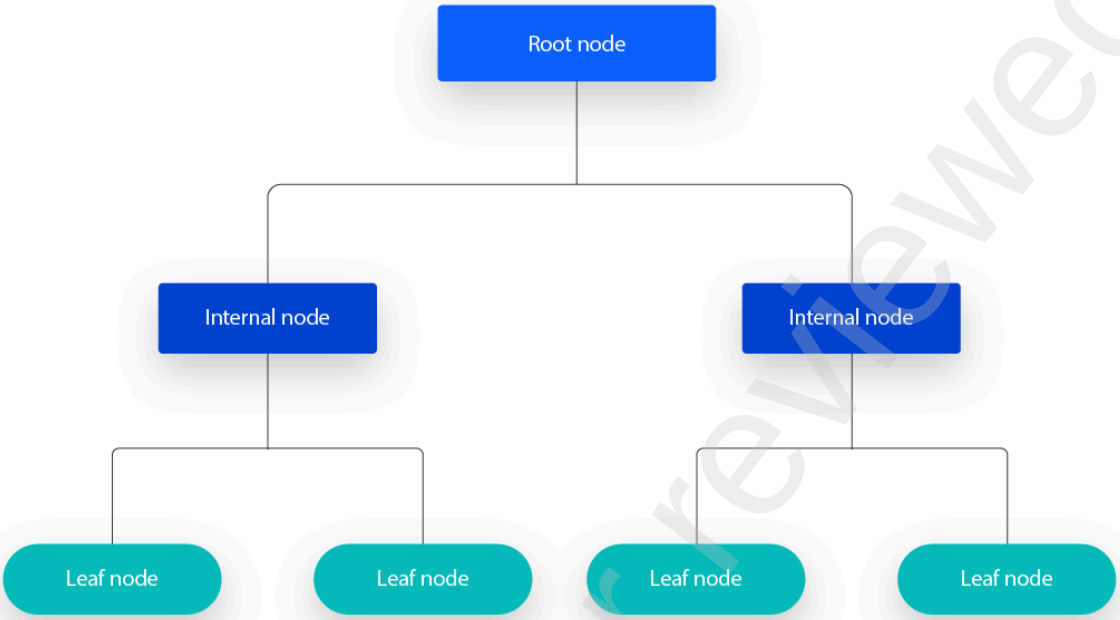

Figure 11. Decision Tree Structure

Delving into the metrics used to assess how well each individual model works, there are several important ways of measurement: accuracy measures how many “correct” identifications the model accomplished (correctly classifying a “normal” patient as “normal” or an “affected” patient as “affected), precision measures the number of true positives divided by total positives, recall accounts for the number of true positives divided by the total number of true positive and false negatives (practically, out of all that were affected, how many did the model accurately predict), and finally F1 score, which is an average of the precision and recall values.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9107142857142857	0.9107142857142857	1.0	0.9532710280373832
Naive Bayes	0.8571428571428571	0.9387755102040817	0.9019607843137255	0.92
K Neighbors	0.8571428571428571	0.9215686274509803	0.9215686274509803	0.9215686274509803
Random Forest	0.9107142857142857	0.9423076923076923	0.9607843137254902	0.9514563106796117

Table 3: Varying Models Applied & Corresponding Metrics

Discussion

Based on the metrics indicated by the prior models, logistic regression and random forest were the most accurate predictors of muscular dystrophy presence with an accuracy of over 90%. Specifically, Random Forest also possessed the highest precision, meaning that it didn't overly assume the presence of muscular dystrophy in its diagnoses. Furthermore, Random Forest held the highest F1 score, meaning that overall the model's performance was extremely good, especially in comparison to many of the other models.

This exceptional performance by Random Forest matches prior research, as Random Forest seems to have the most success and apparent usage in current investigation into the topic of muscular dystrophy. For instance, Random Forest was used for meaningful insights for patients with facioscapulohumeral muscular dystrophy, MRI biomarkers for facioscapulohumeral muscular dystrophy, and to differentiate between varying types of muscular dystrophy. Therefore, this performance by the Random Forest model builds off of the work previously done, while also indicating that Logistic Regression and its binary classification may be effective at distinguishing sheerly between affected and unaffected individuals.

However, some discrepancies and flaws with this architecture and programming structure still remain. For instance, some of the models have the exact same values for accuracy and Logistic Regression even attains a status of 100% recall, which based on the accuracy would be rather unlikely. Also, there may have been some specific issues with the merging of the datasets, as some different types of muscular dystrophy may have had different features entirely. Therefore, that dataset with numerous sub archetypes of muscular dystrophy may have been counterproductive in achieving optimal performance.

Conclusion

In conclusion, muscular dystrophy, by nature of its genetic wrath, requires a solution that deals by looking at genes. Machine learning models offer a time-efficient and precise way to process all the genetic information associated with muscular dystrophy and ensure that a diagnosis is completed with relatively reliable results. Specifically, the Random Forest model seems to be the most accurate model when conducting research on genetic information and its role in the onset of muscular dystrophy.

References

Alfano, Lindsay N, and Tahseen Mozaffar. “Random Forest: Random Results or Meaningful Insights for Patients with Facioscapulohumeral Muscular Dystrophy?” *Brain : A Journal of Neurology*, U.S. National Library of Medicine, 16 Dec. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8677539/.

“Articles.” *Cedars*, www.cedars-sinai.org/health-library/diseases-and-conditions/m/myopathy.html#:~:text=Overview,dysfunction%20of%20the%20muscle%20fibers. Accessed 20 July 2024.

Bakay M, Wang Z, Melcon G, Schiltz L et al. Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. *Brain* 2006 Apr;129(Pt 4):996-1013. PMID: 16478798

Chen , Yunji. “Sigmoid Function.” *Sigmoid Function - an Overview | ScienceDirect Topics*, 2007, www.sciencedirect.com/topics/computer-science/sigmoid-function.

Dadgar S, Wang Z, Johnston H, Kesari A et al. Asynchronous remodeling is a driver of failed regeneration in Duchenne muscular dystrophy. *J Cell Biol* 2014 Oct 13;207(1):139-58. PMID: 25313409

Gopalakrishnan, T., et al. “An Automated Deep Learning Based Muscular Dystrophy Detection and Classification Model.” *Tech Science Press*, Tech Science Press, 3 Nov. 2021, www.techscience.com/cmc/v71n1/45392/html.

Jangra, Ranjeet. “Features and Labels in Ai.” *Medium*, Medium, 27 Jan. 2024, medium.com/@ranjeetjangra/features-and-labels-in-ai-bb66b78a93b8.

LaPelusa, Andrew. “Muscular Dystrophy.” *StatPearls [Internet]*., U.S. National Library of Medicine, 26 Feb. 2024, www.ncbi.nlm.nih.gov/books/NBK560582/.

Lumivero. “Ordinary Least Squares Regression (OLS).” *XLSTAT, Your Data Analysis Solution*, www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols. Accessed 20 July 2024.

Medicine , Johns Hopkins. “Types of Muscular Dystrophy and Neuromuscular Diseases.” *Johns Hopkins Medicine*, 15 July 2024, www.hopkinsmedicine.org/health/conditions-and-diseases/types-of-muscular-dystrophy-and-neuromuscular-diseases#:~:text=Muscular%20dystrophy%20is%20a%20group,increasing%20disability%2C%20and%20possible%20deformity.

Monforte M;Bortolani S;Torchia E;Cristiano L;Laschena F;Tartaglione T;Ricci E;Tasca G; "Diagnostic Magnetic Resonance Imaging Biomarkers for Facioscapulohumeral Muscular Dystrophy Identified by Machine Learning." *Journal of Neurology*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/34486074/. Accessed 20 July 2024.

Pescatori, Mario, et al. "Gene expression profiling in the early phases of DMD: A constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression." *The FASEB Journal*, vol. 21, no. 4, 30 Jan. 2007, pp. 1210–1226, https://doi.org/10.1096/fj.06-7285com.

Pfizer. "Duchenne Muscular Dystrophy." *Pfizer*, www.pfizer.com/disease-and-conditions/duchenne-muscular-dystrophy#:~:text=In%20total%2C%20dystrophy%20disorders%20affect,people%20assigned%20male%20at%20birth.&text=About%20250%2C000%20people%20in%20the,some%20form%20of%20muscular%20dystrophy.&text=Experts%20estimate%20fewer%20than%2050%2C000%20people%20in%20the%20United%20States%20have%20DMD. Accessed 20 July 2024.

Researcher, Vijay Kanade AI, et al. "Everything You Need to Know about Logistic Regression - Spiceworks." *Spiceworks Inc*, 13 May 2024, www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20a%20supervised%20machine%20learning%20algorithm%20that%20accomplishes,1%2C%20or%20true%2Ffalse.

Rish , Irina. "(PDF) an Empirical Study of the Naïve Bayes Classifier." *Researchgate*, 2001, www.researchgate.net/publication/228845263_An_Empirical_Study_of_the_Naive_Bayes_Classifier.

Ross, Nicola, and Sarah Marsh . "Neuromuscular Disorders & Anaesthesia." *WFSA Resource Library*, 5 Sept. 2022, resources.wfsahq.org/atotw/neuromuscular-disorders-anaesthesia/.

"What Is a Decision Tree?" *IBM*, 2 Nov. 2021, www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.