MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

DEPARTMENT OF COMMERCE

SUBJECT CODE: Minor Project – II (MSBA6304)

Classification and Analysis of the type of Spinal Muscular Atrophy

Report Submitted By

Creig Luke Picardo Reg. No: 232626036

Sowmya M Kamath Reg. No: 232626011

Harishankar Arun C Reg. No: 232626033

MSc. Business Analytics

Under the Guidance of

Dr. Mathew Thomas Gil

Assistant Professor Sr Scale

Department of Commerce

Manipal Academy of Higher Education

Manipal, India

November 2024

Declaration

We hereby declare that this project report titled **"Classification and Analysis of the type of Spinal Muscular Atrophy"** submitted byus to Manipal Academy of Higher Education in partial fulfilment of the requirements of the Degree of MSc. Business Analytics is a record of bona fide work carried out by me under the guidance of Dr. Mathew Thomas Gil Department of Commerce, Manipal Academy of HigherEducation, Manipal.

This report is the outcome of our work with Dr. Mathew Thomas Gil and the contents of this report have not previously formed the basis for the award of any Degree, Diploma or such other similar title.

Date: 24-11-2024

Place: Manipal

Creig Luke Picardo (232626036)

Sowmya M Kamath (232626011)

Harishankar Arun C (232626033)

Department of Commerce

MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

CERTIFICATE

This is to certify that this thesis titled " **Classification and Analysis of the type of Spinal Muscular Atrophy**" is submitted by Creig Luke Picardo (Reg. No. 232626036), Sowmya M Kamath (Reg. No. 232626011), Harishankar Arun C (Reg. No. 232626033) to the Department of Commerce, Manipal Academy of Higher Education (MAHE), towards partial fulfilment of the requirements for the award of the degree of M.Sc. Business Analytics. Creig Luke Picardo (Reg. No. 232626036), Sowmya M Kamath (Reg. No. 232626011), Harishankar Arun C (Reg. No. 232626033) have worked under my supervision and guidance. No part of this report has been earlier submitted for the award of any Degree, Diploma, Fellowship or any other similar title or prizes. The work has not been published in any journal or magazine.

Date: 24-11-2024
Place: Manipal

Dr. Mathew Thomas Gil

New Management Block, Near 9th Block, MIT Campus, Manipal 576 104, Karnataka, India• Ph: 91820 2925342 • Fax: 91 820 2925341 • e-mail: doc.office@manipal.edu

ACKNOWLEDGEMENT

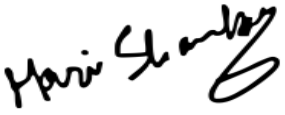
We believe that our project will be complete only after we thank the people who have helped and contributed to making this project successful.

First and foremost, we would like to thank our beloved Head of Department, Dr. Sandeep Shenoy, for giving us an opportunity to carry out our project work at college and providing us with all the facilities needed.

I express my deep sense of gratitude and indebtedness to my guide Dr. Mathew Thomas Gil, Department of Commerce, for his inspiring guidance, constant encouragement, support, and suggestions for improvement during the course for our project.

We are fortunate enough to get constant support and encouragement from all the teaching and non-teaching staff of the Department of Commerce. I also thank the people who were directly or indirectly instrumental in the successful completion of our project.

Last but not the least, I would like to thank all our well-wishers for their unconditional support provided during the data collection period and the rest of our project development.



Creig Luke Picardo
Reg.: 232626036

Sowmya M Kamath
Reg.: 232626011

Harishankar Arun C
Reg.: 232626033

TABLE OF CONTENTS

Contents **Page No.**

LIST OF FIGURES

ABSTRACT

Spinal Muscular Atrophy (SMA) is a genetic neuromuscular disorder, characterized by motor neuron degeneration, which leads to the progressive weakness and atrophy of body muscles. Mutations or deletions of the Survivor Motor Neuron genes in the DNA, SMN1 or/and SMN2 gene primarily cause this disorder. This leads to inadequate production of the SMN proteins, which is indispensable for the sustenance and survival of motor neurons. There exist different types of SMA related to the onset age and progression of symptoms: Type 1, Type 2, Type 3 and Type 4. Advances in genetic testing have improved significantly the rate of diagnosis of SMA at an early stage in life so that interventions may occur appropriately and manage symptoms to improve the life quality of individuals diagnosed with SMA. Over the years, clinical understanding of SMA has changed, with much research currently focused on its genetic basis, clinical manifestations, and therapeutic strategy. Machine learning approaches, especially Supervised and Unsupervised learning models, have been widely used in the setting of SMA for predictive modelling and early detection. These models use genetic data, patient demographics, locomotory data and other clinical features to classify patterns related to SMA onset and severity. Studies have shown that ML algorithms predict the probability of SMA in those carrying SMN1 gene mutations and thereby open promising avenues for enhancing diagnostic accuracy, predominantly in cases presymptomatic. Moreover, monitoring and optimizing SMA treatment protocols using ML models also contribute to personalized management of SMA patients. As this field grows, the adoption of ML models into real practice will change the face of SMA diagnosis and treatment, allowing more precise, time-sensitive interventions in affected patients.

Keywords: Spinal Muscular Atrophy, Genetic, Classification, Machine Learning, detection

CHAPTER 1: INTRODUCTION

Spinal Muscular Atrophy (SMA) refers to a kind of gene inflicted disease caused by the mutation of the Survivor Motor Neuron gene, SMN1 and SMN2, which predisposes patients to the atrophy of the voluntary skeletal muscles due to lower motor neuron degeneration. SMA is an inherited disease which occurs when both parents have a missing part of SMN1/SMN2 Gene. If one parent has the recessive gene and the other parent has a regular gene, the child will not have SMA syndrome, but they might be the carrier of the recessive gene.

SMA is one of the most common genetic pathologies that contributes to the death of infants within their first years of life. Manifests of SMA are different from one individual to the other and are determined by genetic predisposition as well as the age at which a particular person begins to show the symptoms of the disease. In the past, a diagnosis of SMA was primarily based on genetic testing but as there are many forms of the disease with varying rates of progression, the monitoring of the disease and devising appropriate therapeutic intervention still poses serious concerns.

New technology in the form of in-depth machine learning (ML) represents great potential in the area of understanding, diagnosing, and treating SMA. Machine Learning (ML) looks for connections between different clinical data such as genetic markers, patient history, or clinical measurements where one data set on its own when analysed in the traditional way would not make sense.

Machine learning is useful for understanding the likelihood of developing the disease and its progression, typing the patients, finding predictors indicating the severity of the disease, and so on. These insights could also be beneficial in the beginning the treatment at the right time and obtaining the expected outcomes of treatment respect to the disease progress and the patient.

The objective of the research is to correctly predict the type of Spinal Muscular Atrophy based on age, Genetic and locomotory factors with the assistance of Machine Learning modelling. The project explores three Machine Learning models, Random Forest Algorithm (RF), Support Vector Machine (SVM) and Gradient Boosting (XGBoost) algorithms to classify the type of SMA.

OBJECTIVE

1. To classify the type of Spinal Muscular Atrophy of an individual based on the genetic and locomotory factors of the disease.

2. To check which factors are the most important and pivotal and deciding the Type of SMA of an individual.

3. To analyze how each of the genetic and locomotory factors help in influencing the type of SMA an individual has.

CHAPTER 2: LITERATURE REVIEW

Spinal Muscular Atrophy (SMA) research has become increasingly important in understanding the disease and improving patient care. This literature review aims to explore how recent advancements in gene therapy, pharmacological treatments, and early diagnostic strategies impact key aspects of SMA management, including patient outcomes, quality of life, and access to treatment. It also explores how ML Algorithms can be used to improve SMA detection in patients.

Nishio et al. (2024) provided a comprehensive overview of the historical evolution of SMA diagnosis and treatment, examining past techniques and recent innovations while forecasting future trends. Their study highlights how advancements in genetic and molecular diagnostics have reshaped the understanding and management of SMA.

Similarly, **Pankaj et al. (2024)** conducted a review of therapeutic advancements, identifying the latest treatment strategies and the persisting challenges in providing effective care to SMA patients.

Gene-targeting therapies represent a significant breakthrough in SMA treatment. **Haque and Yokota (2024)** explored these therapies' potential to improve SMA outcomes. Their work emphasized the promise of gene-targeting interventions, such as SMN1 gene replacement and SMN2 gene modulation, while also addressing challenges related to their clinical application.

The systemic nature of SMA extends beyond motor neuron degeneration, influencing other biological pathways and healthcare systems. **Lipnick et al. (2024)** utilized insurance claims data to uncover SMA's broader impact on patients and healthcare systems, offering novel insights into the disease's systemic burden.

Additionally, **Tapken et al. (2024)** examined SMA’s molecular complexity, revealing how it affects various biological pathways beyond motor neurons. Their findings underscore the importance of viewing SMA as a multifaceted disorder rather than a localized motor neuron disease.

Sun et al. (2024) employed single-cell RNA sequencing in a severe SMA mouse model to identify dysregulation in spinal cord cell types. This study provided crucial insights into the cellular and molecular mechanisms underlying SMA, paving the way for targeted interventions aimed at restoring cellular homeostasis.

Biomarkers have emerged as critical tools in monitoring SMA progression and treatment efficacy. **Maretina et al. (2024)** explored the relevance of biomarkers in the treatment era, highlighting their potential to refine patient management by tracking disease progression and therapeutic responses.

Understanding the psychosocial dimensions of SMA is vital for holistic patient care. **De Lemus et al. (2024)**, through the PROfuture project, identified key aspects of SMA impacting the quality of life of patients and caregivers. Their qualitative study revealed the profound physical and emotional burdens associated with the disease, emphasizing the need for comprehensive support systems

CHAPTER 3: RESEARCH METHODOLOGY

Conceptual Framework

Factors to consider:

F1: Gender
F2: Diagnosis age
F3: Diagnostic Test Type
F4: Zygosity (Genetic study result)
F5: Mutation Type
F6: SMN2 Copies
F7: Gastrostomy
F8: Tracheostomy
F9: Locomotion

Here's a brief explanation of how each of these factors may influence the type of Spinal Muscular Atrophy (SMA):

1. **Gender (F1):** SMA affects both males and females equally in terms of genetics. However, gender might indirectly influence clinical outcomes, healthcare access, or the progression of certain symptoms, though these are generally minor factors in SMA type classification. The impact of gender on type of SMA is very minimal.

2. **Diagnosis Age (F2):** The age at diagnosis often correlates with the severity and type of SMA. Earlier onset (usually in infancy) is associated with more severe types (e.g., Type 1), while later onset often aligns with milder forms (e.g., Type 3 or Type 4).

3. **Diagnostic Test Type (F3):** Diagnostic methods (genetic testing, enzyme tests, or muscle biopsies) confirm SMA presence and type but don't directly influence it. However, certain tests may be more prevalent for detecting specific mutations associated with different SMA types.

4. **Zygosity (F4) - Genetic Study Result:** Zygosity (whether an individual has one or both defective SMN1 genes) can impact disease severity. Homozygous mutations in the SMN1 gene (two copies with mutations) are typical in more severe SMA types, while heterozygous mutations (one mutated and one normal gene) can sometimes result in milder phenotypes.

5. **Mutation Type (F5):** SMA types are strongly linked to mutations of the SMN1 gene. The specific mutation and its location can influence disease severity, with certain mutations being more common in severe SMA types (Type 1) and others in milder forms (Types 3 and 4).

6. **SMN2 Copies (F6):** The number of SMN2 Gene copies is a major modifier in SMA. More SMN2 copies generally lead to milder SMA types because the SMN2 gene produces some functional SMN protein. Type 1 often has only one or two SMN2 copies, while Types 3 and 4 might have three or more.

7. **Gastrostomy (F7):** Patients requiring a feeding tube (gastrostomy) tend to have severe forms of SMA (e.g., Type 1 or 2) due to compromised muscle function impacting swallowing and breathing.

8. **Tracheostomy (F8):** The need for a tracheostomy usually indicates severe respiratory weakness and is more common in SMA Types 1 and 2. This intervention is less likely in milder types like Types 3 or 4, where respiratory function is better preserved.

9. **Locomotion (F9):** Locomotion plays a very important role in detecting the type of SMA. Individuals with SMA Type 1 have little to no movement of muscles. SMA Type 2 individuals exhibit a little movement and can sit upright. Individuals with SMA Type 3 usually can move short distances with minimal help, whereas Type 4 SMA individuals have minimal muscle weakness and can move about almost freely.

Data Collection

For the study we took secondary data from Harvard Dataverse. The data pertains to the Spinal Muscular Atrophy (SMA) cases in the states of Colombia in South America.

Dataset features: The dataset contains columns that are an interplay of String, ordinal and categorical data.

The columns ‘City’ and ‘State’ are string data. Column ‘Diagnosis Age’ is ordinal data, while the remaining data points are all nominal data, converted to factors.

Methodology

Data Visualization: Data Visualization was done using PowerBI and RStudio. The visualization gives a bird's eye view of the data in hand, considering various factors such as age, Gender, Gene expression, etc.

Random Forest Algorithm: Random Forest (RF) Algorithm is an ensemble ML algorithm that builds decision trees on a random set of features, and the results are combined to improve accuracy and reduce overfitting. The features age, diagnostic test, Genetic study, Mutation Type, SMN2 Copies, Gastrostomy, Tracheostomy, Locomotion were used to train the model, while the factor SMA Type (1, 2, 3, 4) was predicted using the model.

Support Vector Machine (SVM): Support Vector Machine (SVM) is an algorithm used to classify data into different categories. The algorithm aims to find the optimal hyperplane (Support Vector) that maximizes the distance between the different features in the dataset. The features age, diagnostic test, Genetic study, Mutation Type, SMN2 Copies, Gastrostomy, Tracheostomy, Locomotion were used to train the model, while the factor SMA Type (1, 2, 3, 4) was predicted using the model.

Gradient Boosting Algorithm: Gradient Boosting (XGBoost) is an ML algorithm used for regression and classification. It builds an ensemble of weak learners by optimizing a loss function. A provision of Gradient Boosting algorithm is that it can combine multiple weak models (shallow trees) into a singular strong model, with iterative focus on error-correction of previous models. Gradient Descent is used to minimize the loss function of the model. The same features that were used for previous models were used to train the GB model, and the factor Type of SMA (1, 2, 3, 4) was predicted using in-sample test data.

Ethical Considerations

The dataset was Harvard's Dataverse library which is public domain.

Variables used for analysis

1. **Diagnosis Age (F2):** The age at which SMA was diagnosed.

2. **Diagnostic Test Type (F3):** Diagnostic method used for SMA detection; Multiplex ligation-dependent probe amplification (MLPA) or Sequencing

3. **Zygosity (F4) - Genetic Study Result:** Zygosity (whether an individual has one or both defective SMN1 genes).

4. **Mutation Type (F5):** Mutation Types being Point mutation, exon-7 deletion, Deletion of exon 7 to 8, duplication of exon 7 to 8.

5. **SMN2 Copies (F6):** The number of copies of the genetic material SMN2. The copies can range in quantity from 1 to 5.

6. **Gastrostomy (F7):** The factor for patients requiring a feeding tube (gastrostomy).

7. **Tracheostomy (F8):** The factor for patients requiring a breathing tube (tracheostomy).

8. **Locomotion (F9):** The amount of movement the patient with SMA can make: No movement, minimal movement, Sit, Walk, or mild leg weakness.

CHAPTER 4: DATA ANALYSIS AND INTERPRETATION

Data Visualization

Fig 1.1 The number of Survivor Motor Neuron-1 (SMN1) copies for each SMA type

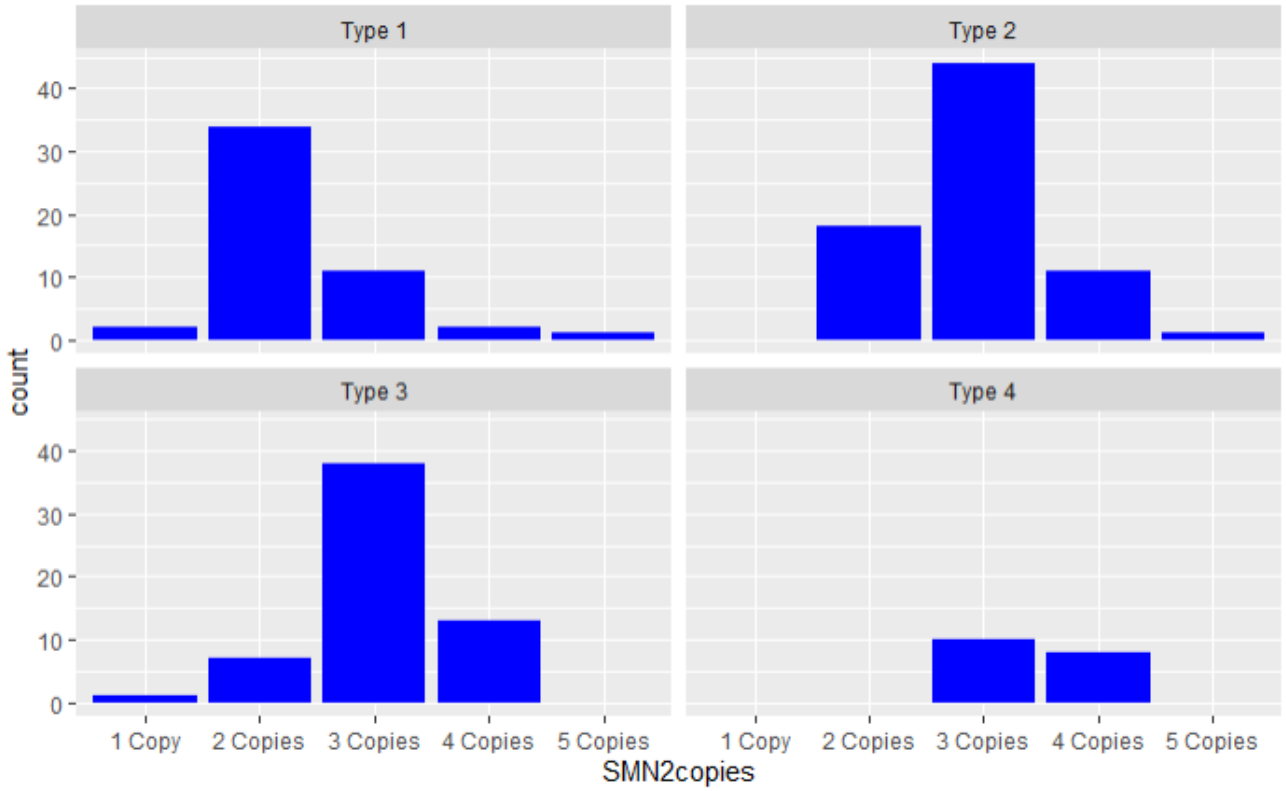


Fig1.2 SMA Types by Locomotion (Movement) of the individual

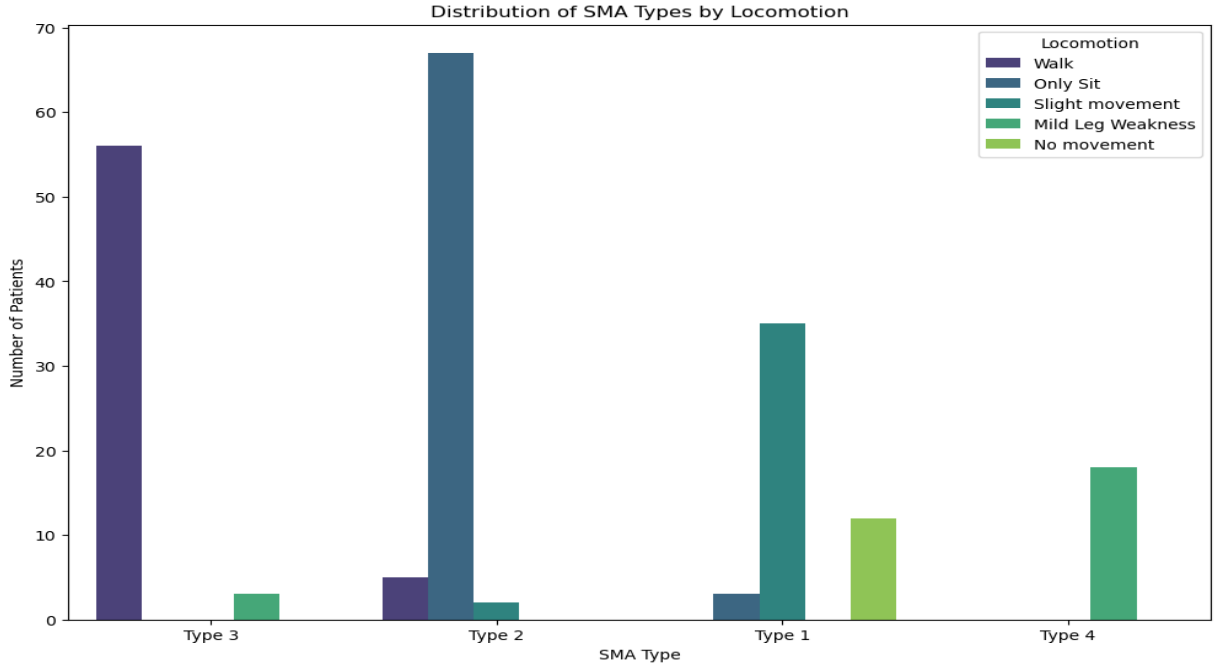


Fig 1.3 Violin Plot for the Patient age filtered by SMA Type

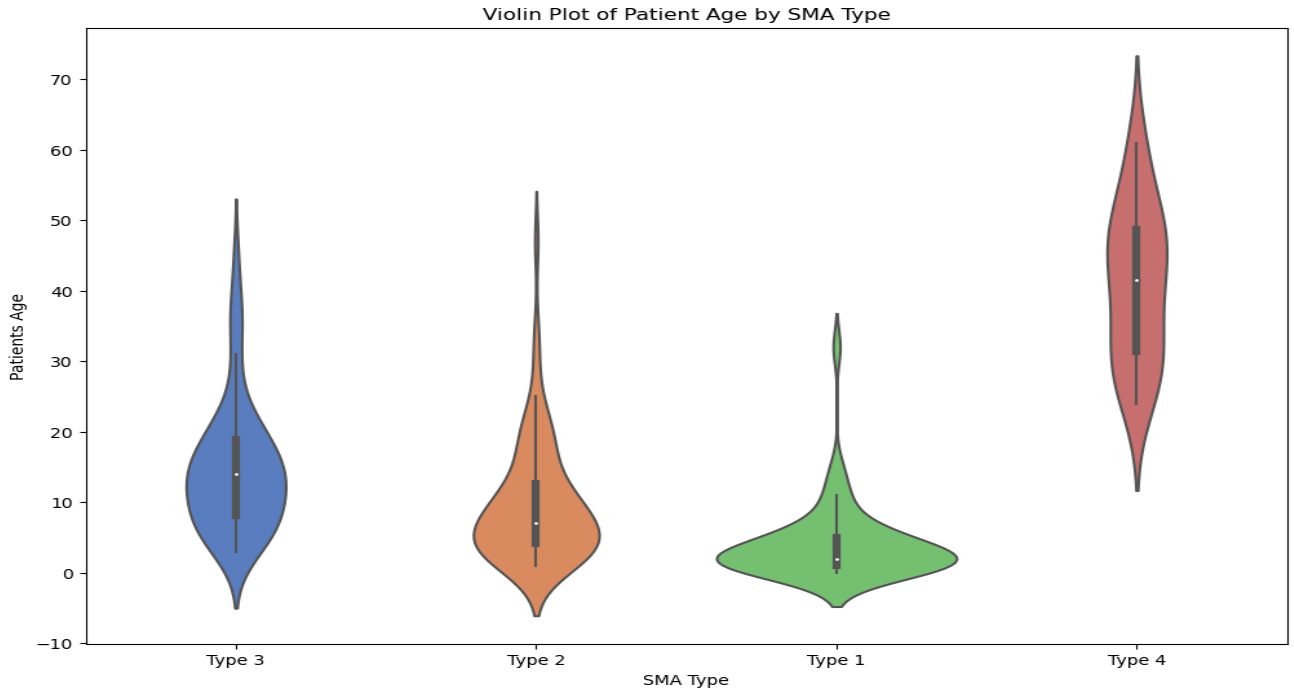
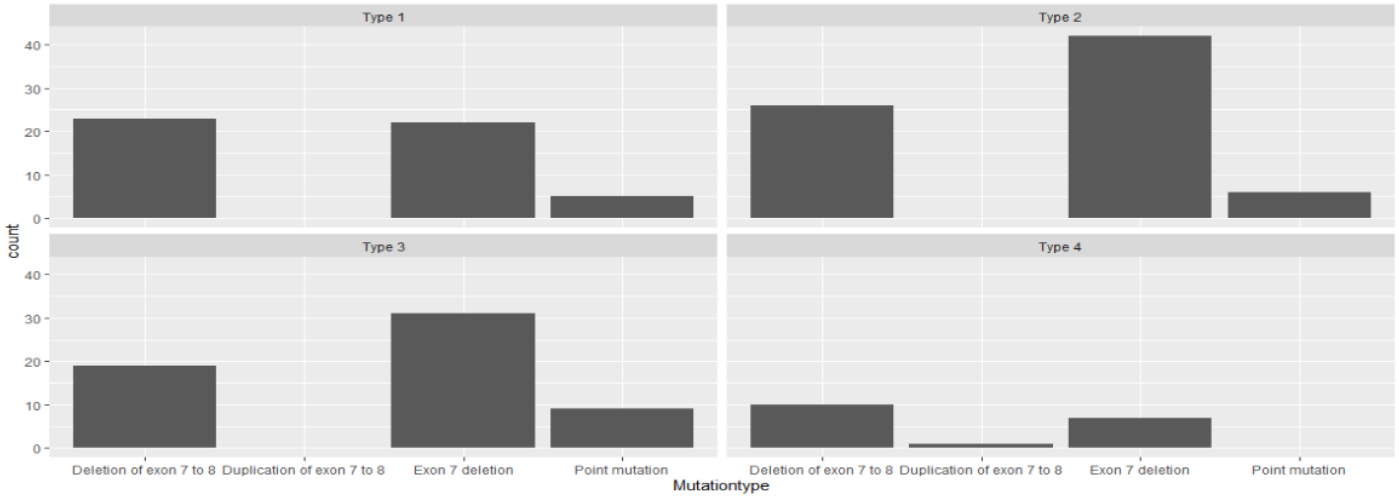


Fig 1.4 Mutation of SMN1 based on the type of SMA



Descriptive Statistics:

Overall Data

The mean age of the individuals is 12.5 years. Majority of individuals having SMA exhibit some form of locomotory movement. Majority of the individuals have more than 2 SMN2 gene Copies. Majority of the individuals do not need a gastrostomy or tracheostomy tube.

Fig 2.1

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Diagnosisage	1	201	12.53	12.72	9	10.14	8.90	1	61
RGender	2	201	1.42	0.50	1	1.40	0.00	1	2
RDiagnosticTest	3	201	1.06	0.25	1	1.00	0.00	1	2
RGeneticStudy	4	201	1.08	0.28	1	1.00	0.00	1	2
RMutationType	5	201	2.30	0.65	2	2.37	0.00	1	4
RSMN2Copies	6	201	2.87	0.74	3	2.84	0.00	1	5
RGastrostomy	7	201	0.09	0.29	0	0.00	0.00	0	1
RTracheostomy	8	201	0.07	0.26	0	0.00	0.00	0	1
RLocomotion	9	201	3.21	1.05	3	3.21	1.48	1	5
SMAtype*	10	201	2.22	0.92	2	2.17	1.48	1	4

SMA Type 1

For SMA Type 1, the mean age is 4 years, and most individuals have very minimal movement (1.82). The mean amount of SMN2 Copies is 2.32. A good 1/3rd of the individuals (0.32) need a gastrostomy tube and 1/4th of the individuals need a tracheostomy tube.

Fig 2.2

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>
Diagnosisage	1	50	4.02	5.23	2
RGender	2	50	1.38	0.49	1
RDiagnosticTest	3	50	1.08	0.27	1
RGeneticStudy	4	50	1.10	0.30	1
RMutationType	5	50	2.36	0.66	2
RSMN2Copies	6	50	2.32	0.71	2
RGastrostomy	7	50	0.32	0.47	0
RTracheostomy	8	50	0.24	0.43	0
RLocomotion	9	50	1.82	0.52	2

SMA Type 2

For SMA Type 2, the mean age is 9.7 years, and most individuals have very decent leg and body movement (3.04). The mean amount of SMN2 Copies is 2.93. Only a small number of individuals need a gastrostomy and tracheostomy tube.

Fig 2.3

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>
Diagnosisage	1	74	9.68	8.40	7
RGender	2	74	1.42	0.50	1
RDiagnosticTest	3	74	1.04	0.20	1
RGeneticStudy	4	74	1.05	0.23	1
RMutationType	5	74	2.27	0.60	2
RSMN2Copies	6	74	2.93	0.67	3
RGastrostomy	7	74	0.03	0.16	0
RTracheostomy	8	74	0.03	0.16	0
RLocomotion	9	74	3.04	0.31	3

SMA Type 3

For SMA Type 3, the mean age is 14.66 years, and most individuals have very decent leg and body movement (4.05). The mean amount of SMN2 Copies is 3.07. Negligible individuals need a gastrostomy tube, and no individual requires a tracheostomy tube for SMA Type 3.

Fig 2.4

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>
Diagnosisage	1	59	14.66	9.08	14
RGender	2	59	1.46	0.50	1
RDiagnosticTest	3	59	1.10	0.30	1
RGeneticStudy	4	59	1.14	0.35	1
RMutationType	5	59	2.17	0.67	2
RSMN2Copies	6	59	3.07	0.64	3
RGastrostomy	7	59	0.02	0.13	0
RTracheostomy	8	59	0.00	0.00	0
RLocomotion	9	59	4.05	0.22	4

SMA Type 4

For SMA Type 4, 40.9 years is the mean age, and all individuals have good leg and body movement, with only mild muscle weakness (5). The mean amount of SMN2 Copies is 3.44. No individual of SMA Type 4 needs a gastrostomy or tracheostomy tube.

Fig 2.5

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>
Diagnosisage	1	18	40.94	11.00	41.5
RGender	2	18	1.44	0.51	1.0
RDiagnosticTest	3	18	1.00	0.00	1.0
RGeneticStudy	4	18	1.00	0.00	1.0
RMutationType	5	18	2.67	0.59	3.0
RSMN2Copies	6	18	3.44	0.51	3.0
RGastrostomy	7	18	0.00	0.00	0.0
RTracheostomy	8	18	0.00	0.00	0.0
RLocomotion	9	18	5.00	0.00	5.0

Random Forest Model

Model Summary

The Random Forest Model summary. Variables Diagnosis age, Diagnostic test, Genetic Study, Mutation Type, No. of SMN2 Copies, Gastrostomy, Tracheostomy and locomotion were taken into consideration to classify SMA Type. The number of trees in the forest are 100, with 2 variables at each split.

Fig3.1

```
Call:
 randomForest(formula = SMAtype ~ Diagnosisage + RDiagnosticTest +      RGeneticStudy +
RMutationType + RSMN2Copies + RGastrostomy +      RTracheostomy + RLocomotion, data = smadata,
ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 2

        OOB estimate of  error rate: 7.96%
Confusion matrix:
       Type 1 Type 2 Type 3 Type 4 class.error
Type 1     45      5      0      0  0.10000000
Type 2      2     67      5      0  0.09459459
Type 3      0      2     57      0  0.03389831
Type 4      0      0      2     16  0.11111111
```

Fig 3.2. Gini Importance table for each factor

Variables RLocomotion and Diagnosis Age have the highest Gini importance with values 66.96 and 23.26 respectively.

	MeanDecreaseGini
Diagnosisage	23.3679785
RDiagnosticTest	0.5870715
RGeneticStudy	0.5359319
RMutationType	2.5860631
RSMN2Copies	6.1850588
RGastrostomy	4.4900717
RTracheostomy	3.1173899
RLocomotion	66.9608760

Fig* 3.3 *Plotting the importance of each factor variable

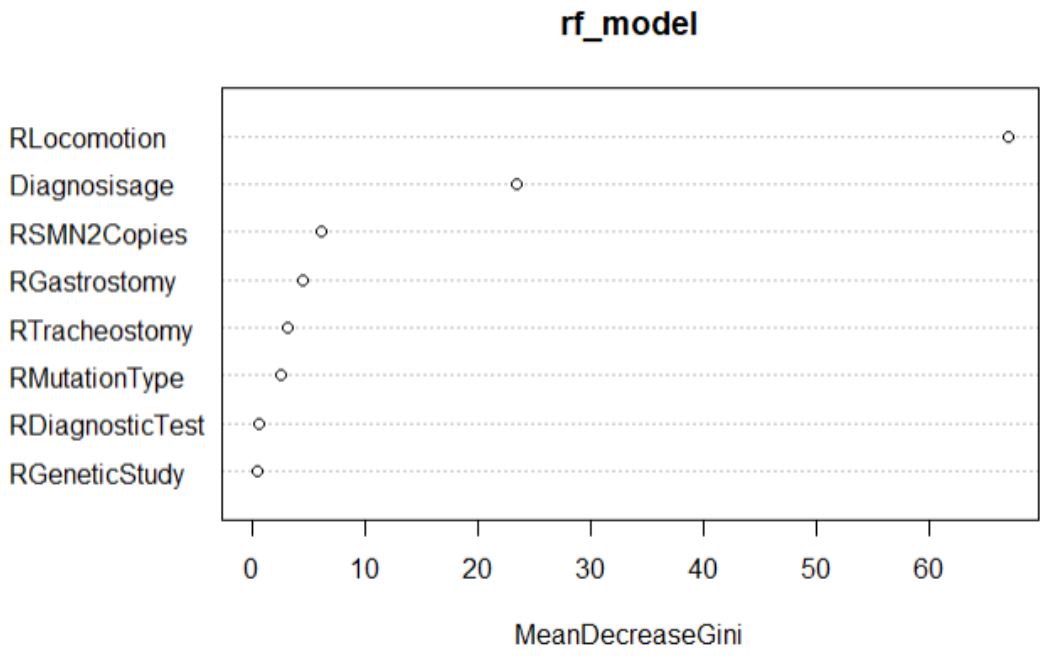


Fig 3.4 Confusion Matrix

38 out of the 40 observations are correctly predicted by the model.

```
          Actual
Predicted Type 1 Type 2 Type 3 Type 4
   Type 1      14       0       0       0
   Type 2       1       3       0       0
   Type 3       0       1       9       0
   Type 4       0       0       0      12
```

Fig 3.5 Accuracy of model is 95%

```
[1] "Accuracy:  0.95"
```

Fig 3.6 OOB Error rate

```
[1] "OOB Error Rate:  0.0796"
```

Support Vector Machine (SVM)

The Support Vector Machine summary. Variables Diagnosis age, Diagnostic test, Genetic Study, Mutation Type, No. of SMN2 Copies, Gastrostomy, Tracheostomy and locomotion were taken into consideration to classify SMA Type with the help pf Support Vector Boundaries. There are 53 support vectors to classify the model.

Fig 4.1

```
Call:
svm(formula = SMAtype ~ Diagnosisage + RDiagnosticTest + RGeneticStudy +
    RMutationType + RSMN2Copies + RGastrostomy + RTracheostomy + RLocomotion,
    data = smadata, kernel = "linear", cost = 1, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  53

 ( 19 20 9 5 )


Number of Classes:  4

Levels:
 Type 1 Type 2 Type 3 Type 4
```

***Fig* 4.2 Confusion Matrix**

The model correctly predicts 191 out of the 201 test cases. Model accuracy is 95.02%

```
pred      Type 1 Type 2 Type 3 Type 4
  Type 1      47      2      0      0
  Type 2       3     67      0      0
  Type 3       0      5     59      0
  Type 4       0      0      0     18
[1] "Accuracy:  95.02 %"
```

***Fig* 4.3 The 2-D visualization of SVM Decision Boundaries**

There is a clear boundary with support vectors to distinguish the 4 classes. Features used for visualization: Age and Locomotion

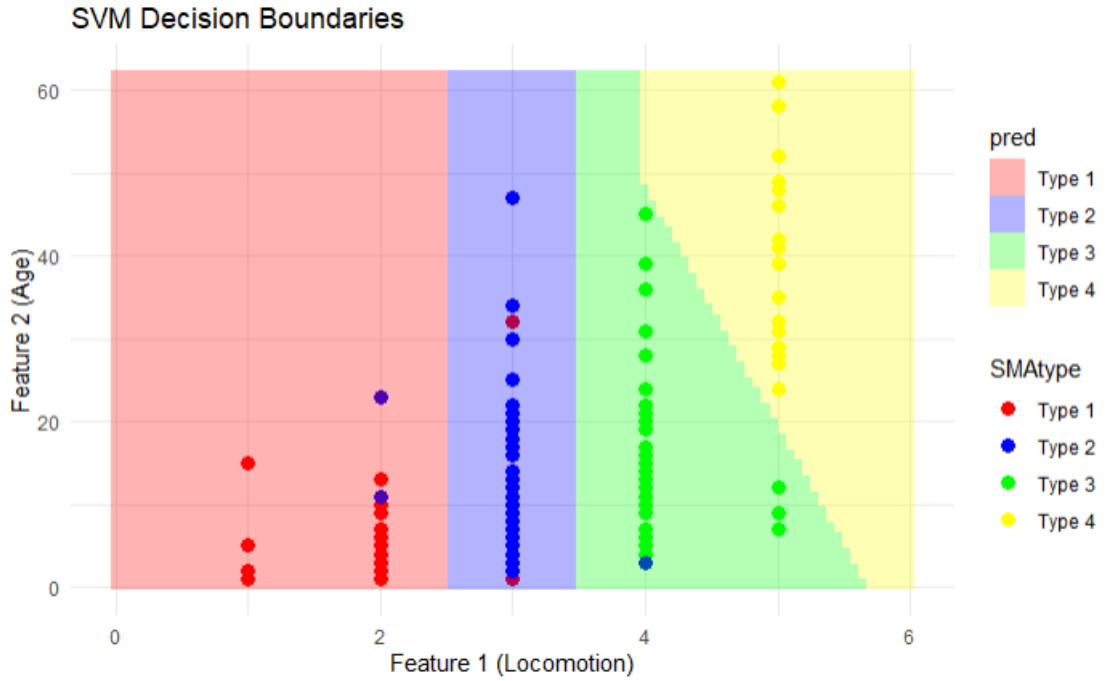
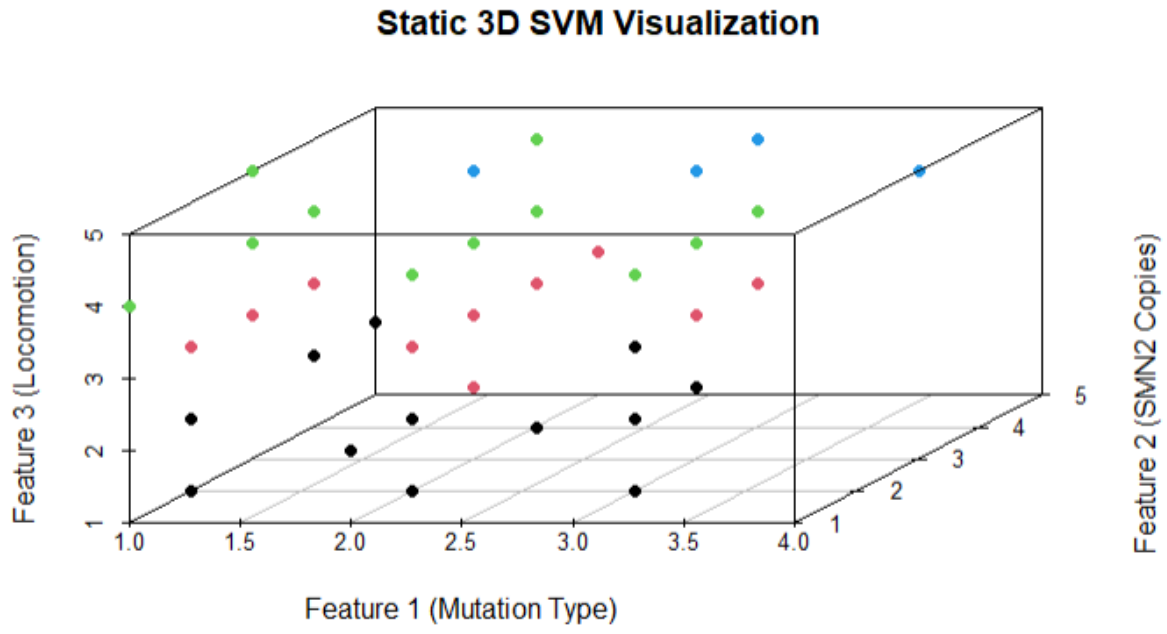


Fig 4.4 3-D Visualization of the SVM Model

Visualizing the Support Vectors using 3 factors
Features used: Locomotion, Mutation Type and Number of SMN2 Copies



Gradient Boosting Model (GBM)

Summary for the Gradient Boosting Model. Variables Diagnosis age, Diagnostic test, Genetic Study, Mutation Type, Gender, No. of SMN2 Copies, Gastrostomy, Tracheostomy and locomotion were taken into consideration to classify SMA Type. 100 iterations were performed.

Fig 5.1

```
call:
  xgb.train(params = params, data = dtrain, nrounds = nrounds,
    watchlist = watchlist, verbose = verbose, print_every_n =
print_every_n,
    early_stopping_rounds = early_stopping_rounds, maximize =
maximize,
    save_period = save_period, save_name = save_name, xgb_model =
xgb_model,
    callbacks = callbacks, max_depth = 4, eta = 0.1, objective =
"multi:softprob",
    num_class = 4)
params (as set within xgb.train):
  max_depth = "4", eta = "0.1", objective = "multi:softprob",
num_class = "4", validate_parameters = "TRUE"
xgb.attributes:
  niter
callbacks:
  cb.evaluation.log()
# of features: 9
niter: 100
nfeatures : 9
evaluation_log:
```

***Fig* 5.2 Loss value with each subsequent iteration**

iter <dbl>	train_mlogloss <dbl>
1	1.22426280
2	1.09082166
3	0.97852896
4	0.88263740
5	0.79949942
6	0.72708581
7	0.66404703
8	0.60812721
9	0.55851912
10	0.51439854

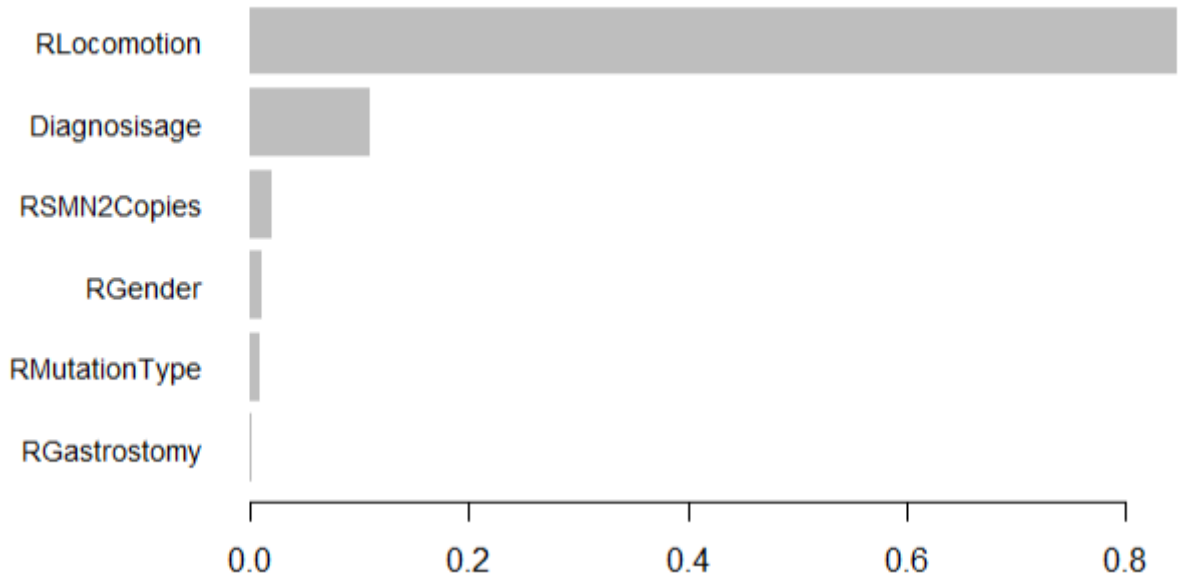
***Fig 5.3* Accuracy**

Model Accuracy: 93.44%

```
[1] "Accuracy: 93.44 %"
```

***Fig 5.4* Feature Importance Plot**

Most important features: Locomotion, Diagnosis age and SMN2 Copies



CHAPTER 5: FINDINGS AND CONCLUSIONS

Ensemble classification was done for the data using 3 models, namely Random Forest Model, Support Vector Machine and Gradient Boosting (XGBoost). The following are the findings from the three Machine Learning models:

1. Random Forest Model

- **Features Used:** Diagnosis Age, Diagnostic Test, Genetic Study, Mutation Type, Number of SMN2 Copies, Gastrostomy, Tracheostomy, and Locomotion.

- **Key Insights:**
  - **Feature Importance:** Locomotion (Gini importance: 66.96) and Diagnosis Age (Gini importance: 23.26) are the most influential factors in classifying SMA types.
  - **Accuracy:** 95%.
  - **Performance:** Correctly classified 38 out of 40 observations.
  - **OOB Error:** Error stabilizes with 100 trees, indicating a well-trained model.

High accuracy and robust performance across SMA types.

2. Support Vector Machine (SVM)

- **Features Used:** Diagnosis Age, Diagnostic Test, Genetic Study, Mutation Type, Number of SMN2 Copies, Gastrostomy, Tracheostomy, and Locomotion.

- **Key Insights:**
  - **Support Vectors:** 53 support vectors used to create decision boundaries.
  - **Accuracy:** 95.02%.
  - **Performance:** Correctly classified 191 out of 201 test cases.
  - **Visualization:**
    - **2D Boundaries:** Clear distinction among SMA classes using Locomotion and Diagnosis Age.
    - **3D Visualization:** Effective separation using Locomotion, Mutation Type, and SMN2 Copies.

The model is excellent for identifying complex class separations, but computationally complex.

3. Gradient Boosting Model (GBM)

- **Features Used:** Diagnosis Age, Diagnostic Test, Genetic Study, Mutation Type, Gender, Number of SMN2 Copies, Gastrostomy, Tracheostomy, and Locomotion.

- **Key Insights:**
  - **Feature Importance:** Locomotion, Diagnosis Age, and SMN2 Copies are the most important factors.
  - **Accuracy:** 93.44%.
  - **Performance:** 100 iterations reduced the loss function effectively.
- **Strengths:** Gradual improvement with boosting iterations, suitable for non-linear relationships.
- **Weaknesses:** Slightly lower accuracy when pitted against Random Forest and SVM.

Overall Conclusion

- **Best Model:** Both Random Forest and SVM perform equally well in terms of accuracy (~95%). However:

  - **Random Forest** is better suited for interpretability and robustness to noisy data.
  - **SVM** excels in cases requiring precise decision boundaries, especially with complex feature interactions.

- **Gradient Boosting Model** is effective but slightly underperformed (93.44% accuracy) compared to the other two.

Most important Features for SMA Prediction

- **Locomotion:** Most impactful across all models, indicating a significant role in distinguishing SMA types.

- **Diagnosis Age:** Second most important, emphasizing the criticality of early diagnosis.

- **Number of SMN2 Copies:** Frequently ranks high, reflecting its genetic relevance to SMA type determination.

- **Mutation Type:** Contributed significantly in SVM and GBM visualizations, reinforcing its genetic influence.
  If interpretability is a priority, Random Forest is the best choice. For precise classification with complex relationships, SVM is ideal.

LIMITATIONS AND FURTHER STUDY

The analysis was done on a dataset with data of 201 patients from South America. The accuracy of the models, especially Support Vector Machine (SVM) can be improved by increasing the number of patients datapoints from a diverse geography and demographics.

Other gene factors, such as the expression of IGHMBP2 gene, MORC2 gene, UBA1 gene, DYNC1H1 gene, BICD2 gene, TRPV4 gene in SMA individuals can be explored and data collected to be used for Model building and classification.

REFERENCES

1. ShiNishio, H., Tabe Eko Niba, E., Saito, T., Okamoto, K., Takeshima, Y., & Awano, H. (2024). Spinal Muscular Atrophy: The Past, Present, and Future of Diagnosis and Treatment. Journal of Neuromuscular Disorders. *researchgate.com*.

2. Haque, U. S., & Yokota, T. (2024). Recent Progress in Gene-Targeting Therapies for Spinal Muscular Atrophy: Promises and Challenges. Gene Therapy and Genomics. *researchgate.com*.

3. Lipnick, S. L., Agniel, D. M., Aggarwal, R., Makhortova, N. R., Finlayson, S. G., Brocato, A., Palmer, N., Darras, B. T., Kohane, I., & Rubin, L. L. (2024). Systemic nature of spinal muscular atrophy revealed by studying insurance claims. Healthcare Economics and Outcomes Research. *sciencedirect.com*.

4. Sun, J., Qiu, J., Yang, Q., Ju, Q., Qu, R., Wang, X., Wu, L., & Xing, L. (2024). Single-cell RNA sequencing reveals dysregulation of spinal cord cell types in a severe spinal muscular atrophy mouse model. Cellular and Molecular Neuroscience. *researchgate.com*.

5. Tapken, I., Schweitzer, T., Paganin, M., Schüning, T., Detering, N. T., Sharma, G., Niesert, M., Saffari, A., Kuhn, D., Glynn, A., et al. (2024). The systemic complexity of a monogenic disease: The molecular network of spinal muscular atrophy. Molecular Biology and Genetics. *emerald.com*.

6. Bagga, P., Singh, S., Ram, G., Subham, K., & Singh, A. (2024). Diving into progress: A review on current therapeutic advancements in spinal muscular atrophy. Current Therapeutic Advances in SMA. *researchgate.com.*

7. Mencía de Lemus, M., Cattinari, M. G., Pascual, S. I., Medina, J., García, M., Magallón, A., Dumont, M., & Rebollo, P. (2024). Identification of the most relevant aspects of spinal muscular atrophy (SMA) with impact on the quality of life of SMA patients and their caregivers: The PROfuture project, a qualitative study. Quality of Life in Neurological Disorders. *Sciencedirect.com.*

8. Maretina, M., Koroleva, V., Shchugareva, L., Glotov, A., & Kiselev, A. (2024). The relevance of spinal muscular atrophy biomarkers in the treatment era. Biomarkers and Clinical Applications. *Sciencedirect.com.*

ORIGINALITY REPORT

CREIG LUKE PICARDO - 232626036

ORIGINALITY REPORT

5%	3%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	1%
2	Вукојичић, Александра. "Complement and Microglia Mediate Sensory-Motor Synaptic Loss in Normal Development and in Spinal Muscular Atrophy", University of Belgrade (Serbia), 2024 Publication	<1%
3	Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advances in Networks, Intelligence and Computing - International Conference on Networks, Intelligence and Computing (ICONIC-2023)", CRC Press, 2024 Publication	<1%
4	repositori.udl.cat Internet Source	<1%
5	etd.uwc.ac.za	