CERTIFICATE

This is to certify that the thesis, entitled **"IDENTIFICATION OF RARE GENETIC MUSCULAR DYSTROPHY FROM GENE SEQUENCES AND MUTATION BASED FEATURES THROUGH SHALLOW AND DEEP LEARNING"** submitted to Bharathiar University, in partial fulfillment of the requirements for the award of the Degree of **Doctor of Philosophy in Computer Science** is a record of original research work done by **Mrs. Sathyavikasini. K (Register No. 2013R405)** during the period of September 2013 to August 2017 of her research in the Department of Computer Science at **PSGR Krishnammal College for Women, Coimbatore** under my supervision and guidance and the thesis has not formed the basis for the award of any Degree/Diploma/ Associateship / Fellowship or other similar title to any candidate of any University.

M.S. Vijaya 28/8/17

Countersigned

Signature of the Guide

Dr. M. S. VIJAYA, M.Sc.,M.Phil.,Ph.D.
Associate Professor and Head
Department of Computer Science (PG)
P.S.G.R. Krishnammal College for Women
COIMBATORE - 641 004.

M.S. Vijaya 28/8/17

Head of the Department

Dr. M. S. VIJAYA, M.Sc.,M.Phil.,Ph.D.
Associate Professor and Head
Department of Computer Science (PG)
P.S.G.R. Krishnammal College for Women
COIMBATORE - 641 004.

Lalitha

Principal

PRINCIPAL
PSGR KRISHNAMMAL COLLEGE FOR WOMEN
COIMBATORE - 641 004.

DECLARATION

I, **Sathyavikasini. K (Register No. 2013R405)** hereby declare that the thesis, entitled **"IDENTIFICATION OF RARE GENETIC MUSCULAR DYSTROPHY FROM GENE SEQUENCES AND MUTATION BASED FEATURES THROUGH SHALLOW AND DEEP LEARNING"**, submitted to Bharathiar University, in partial fulfillment of the requirements for the award of the **Degree of Doctor of Philosophy in Computer Science**, is a record of original and independent research done by me during the period of September 2013 to August 2017 under the Supervision and Guidance of **Dr. (Mrs.) M. S. Vijaya M.Sc., M.Phil., Ph.D,** Associate Professor and Head, Department of Computer Science at PSGR Krishnammal College for Women, Coimbatore and it has not been formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or other similar title to any candidate of any University.

Signature of the Candidate

GENUINENESS OF PUBLICATION

GENINUNENESS OF PUBLICATION

This is to certify that the PhD candidate **Mrs. Sathyavikasini. K (Register No. 2013R405)** working under my supervision has published research articles in the following refereed journals.

1. "Muscular Dystrophy Disease Classification Using Relative Synonymous Codon Usage", International Journal of Machine Learning and Computing vol.6, no. 2, ISSN- 2010-3700, pp. 139-144, 2016.
2. "Identification of Rare Genetic Disorder from Single Nucleotide Variants Using Supervised Learning Technique", International Journal of Control Theory and Applications, Vol.9, no.34, pp. 801-810, 2016 (**Scopus indexed**).
3. "Shallow Learning model for diagnosing neuromuscular disorder from splicing variants", World Journal of Engineering, Vol. 14, no. 4, pp.329-336, 2017 (**Scopus Indexed, ISI indexed**).

The contents of the publication incorporates part of the results in his/her thesis.

M.S. Vijaya 28/8/17

Research Supervisor

Dr. M. S. VIJAYA, M.Sc.,M.Phil.,Ph.D.
Associate Professor and Head
Department of Computer Science (PG)
P.S.G.R. Krishnammal College for Women
COIMBATORE - 641 004.

Countersigned

M.S. Vijaya 28/8/17

Head of the Department

Dr. M. S. VIJAYA, M.Sc.,M.Phil.,Ph.D.
Associate Professor and Head
Department of Computer Science (PG)
P.S.G.R. Krishnammal College for Women
COIMBATORE - 641 004.

Nirmala

Principal

PRINCIPAL
PSGR KRISHNAMMAL COLLEGE FOR WOMEN
COIMBATORE - 641 004.

பாரதியார் பல்கலைக்கழகம்

BHARATHIAR UNIVERSITY

COIMBATORE - 641 046, TAMILNADU, INDIA.

State University | Re-accredited with "A" Grade by NAAC | Ranked 14th among Indian Universities by MHRD-NIRF

CERTIFICATE OF PLAGIARISM CHECK

1	Name of the Research Scholar	SATHYAVIKASINI.K
2	Course of study	M.Phil., / Ph.D.,
3	Title of the Thesis / Dissertation	IDENTIFICATION OF RARE GENETIC MUSCULAR DYSTROPHY FROM GENE SEQUENCES AND MUTATION BASED FEATURES THROUGH SHALLOW AND DEEP LEARNING
4	Name of the Supervisor	Dr. M.S. VIJAYA
5	Department / Institution/ Research Centre	DEPARTMENT OF COMPUTER SCIENCE, PSGR KRISHNAMMAL COLLEGE FOR WOMEN
6	% of Similarity of content Identified	07 %
7	Acceptable Maximum Limit	30 %
8	Software Used	URKUND
9	Date of verification	22/ 8 / 2017

Report on plagiarism check, items with % of similarity is attached

M.S. Vijaya

Signature of the Supervisor

(Seal)

Dr. M. S. VIJAYA, M.Sc.,M.Phil.,Ph.D.
Associate Professor and Head
Department of Computer Science (PG)
P.S.G.R. Krishnammal College for Women
COIMBATORE - 641 004.

K. Sathya

Signature of the Researcher

M.S. Vijaya

R&D Director (BU) / Head of the Department

(Seal)

Dr. M. S. VIJAYA, M.Sc.,M.Phil.,Ph.D.
Associate Professor and Head
Department of Computer Science (PG)
P.S.G.R. Krishnammal College for Women
COIMBATORE - 641 004.

University Librarian (BU)

University Librarian
Arignar Anna Central Library
Bharathiar University
641 046

Research Coordinator (BU)

ACKNOWLEGEMENT

I'm blessed and I thank God for every day for everything that happens for me.

For many doctoral students, dissertation process is a long and solitary path to pursue. For my part, it has indeed been challenging but never lonely, thanks to the people who have supported me all along. This is a great milestone in my career and I would like to thank everyone who has helped me to achieve this.

I sincerely express my deep sense of gratitude to **Smt.R.Nandhini**, Chair person, PSGR Krishnammal College for Women for providing the opportunity to pursue the research.

I am indebted to **Dr. (Mrs). N.Yeshodha Devi M.Com., Ph.D,** Secretary, PSGR Krishnammal College for Women for her warm support and inspiration.

I convey my sincere gratitude to **Dr. (Mrs). S. Nirmala MBA., M.Phil., Ph.D,** Principal, PSGR Krishnammal College for Women for all the amenities provided for the conduct of study.

I would like to record here my deep indebtedness and heartfelt gratitude to my research supervisor and guide **Dr. (Mrs). M.S.Vijaya M.Sc., M.Phil., Ph.D,** Associate Professor, Computer Science department, PSGR Krishnammal College for Women for her continuous support, encouragement and motivation, during my research work and for standing by me throughout the period of study without which I would not have completed the research on time. She has always had the ability to see the big depiction and to keep me focused towards the goal. She bestowed endless support and encouragement, which are awfully essential for any doctoral student. No matter how busy she had been, she would always find time to offer counsel for my doubts. I do not think that I could have had a better supervisor than her for my thesis.

I would like to extend my thanks to all the faculty members, lab assistants, librarians and

supporting staff of the Department of Computer Science, PSGR Krishnammal College for Women for their help and encouragement during my course of research.

I am very much thankful to my parents **Mr. K. Kalimuthu** and **Mrs. K. Saraswathy** who encouraged and motivated me towards success through their sound faith and silent prayers. Without them this research work would never come into existence.

I owe my warmest thanks to my father-in-law **Mr. N. Ramalingam** and my mother-in-law **Mrs. R. Saraswathy** for their love and untiring support throughout my life.

This thesis would not have seen the light of day but for the utmost care, persistent motivation and fillip provided in ample measure by my husband, **Mr. R. Muthumanickam** for his positive attitude, which has taken me to great heights. I thank my daughter **Ms. M. Nikarika** who is the best gift I could ever have, for her smile encourages me to efficiently overcome the difficulties encountered in my research. Her love, support and patience has given me a lot of strength and willpower, throughout this arduous journey of research.

I want to thank my family members for their immense love and thoughtful encouragement at home. There are many other good friends and well-wishers who have rendered me help in great or small measure in the successful completion of this project. I am great full to all of them.

If by mistake I have failed to thank anyone, who has slipped from my memory, I offer humble apology and render heartfelt thanks to everyone.

SATHYAVIKASINI. K

List of Figures

1.1	Double Helix Structure of a DNA	7
1.2	Structure of a Gene	8
1.3	Codon Table	12
2.1	Example Decision Tree	40
2.2	Biological Neuron	45
2.3	A Perceptron	45
2.4	Linear Support Vector Machine	50
2.5	Data which require Linear SVM Formulation with Soft Margin	51
2.6	Data that require Non Linear Classifier	52
2.7	Non-linear Mapping into Feature Space	53
2.8	Modeling Techniques of Ensemble Classifier	58
2.9	Multilayer Search Method Schematic Diagram	63
2.10	The Framework of Dynamic Selection and Circulating Combination	64
3.1	Artificial Neuron	73
3.2	Architecture of Deep Neural Network Par diagram	76
3.3	Implementing Stacked autoencoder Phase-I	83
3.4	Implementing Stacked autoencoder Phase-II	84
3.5	Implementing Stacked autoencoder Phase-III	85
3.6	Implementing Stacked autoencoder Phase-IV	85
4.1	HGMD Information for DMD Gene	92

4.2	Sample Output of Generated Mutated Gene Sequence	94
4.3	Disease Identification Model	95
4.4	Process flow of Proposed Muscular Dystrophy Disease Identification Model	98
5.1	Prediction Accuracy of the Classifiers (Non – synonymous mutations)	113
5.2	Precision-Recall Curve of SVM Classifier	114
5.3	ROC Curve of SVM Classifier	114
5.4	Prediction Accuracy of the Classifiers (Synonymous mutations)	121
5.5	Precision-Recall Curve for Decision Tree Classifier	122
5.6	ROC Curve for Decision Tree Classifier	123
5.7	Predictive Accuracy of the Classifiers (Insertion, Deletion and Duplication mutations)	132
5.8	Precision-Recall Curve of SVM Classifier	133
5.9	ROC Curve of SVM Classifier	133
5.10	Predictive Accuracy of the Classifiers(Splicing Mutations)	140
5.11	Precision-Recall Curve of SVM Classifier	141
5.12	ROC Curve of SVM Classifier	142
5.13	Predictive Performance of the Classifiers (Pooled Features)	147
5.14	Predictive Accuracy of the Classifiers (Pooled Features)	147
5.15	Predictive Accuracy of the Classifiers (Pooled Features after feature selection)	148
5.16	Comparative study of Prediction Accuracy Before and After Feature Selection	149
5.17	Precision-Recall curve of SVM Classifier	150
5.18	ROC Curve of SVM Classifier	151
6.1	Prediction Accuracy of LibD3C Classifier (Non Synonymous mutation)	154

6.2	Prediction Accuracy of LibD3C Classifier (Synonymous mutation)	156
6.3	Prediction Accuracy of LibD3C Classifier (Insertion, Deletion and Duplication mutations)	157
6.4	Prediction Accuracy of LibD3C Classifier (Splicing Mutations)	159
6.5	Prediction Accuracy of the LibD3C Classifiers (AGM Mutation)	160
6.6	Performance Comparison of LibD3C with Supervised Models	162
7.1	Translating Nucleotide into Numerical Values	167
7.2	Performance of Classifiers with Tensorflow Deep Neural Network	169
7.3	Performance of Classifiers in Average Training loss with Tensorflow Deep Neural Network	170
7.4	Average Training loss of the Classifier with Nucleotide Mapping	170
7.5	Translating Codon Sequences into Numerical Array	173
7.6	Performance of Deepnet using Codon Mapping	175
7.7	Performance of DNN with Average Training loss using Codon Mapping	175
7.8	Average Training loss of the Classifiers with Codon Mapping	176
7.9	Comparison on Predictive Accuracy in Nucleotide Mapping	178
7.10	Performance of Evaluation Measures in Nucleotide Mapping	179
7.11	Comparison on Predictive Accuracy in Codon Mapping	180
7.12	Performance of Evaluation Measures in Codon Mapping	180

List of Tables

1.1	Muscular Dystrophy Disease Types	21
1.2	Summary of the Existing Work	29
4.1	Genes associated with Different Types of Muscular Dystrophy	90
4.2	Definitions of HGMD Classifications	91
4.3	Training Datasets	97
5.1	Features and their Descriptions	107
5.2	Predictive Performance of the Classifiers (Non – Synonymous Mutations)	112
5.3	Performance Evaluation of the Classifiers (Non-synonymous mutations)	113
5.4	RSCU Values for 59 Codons	117
5.5	Predictive Performance of the Classifiers (Synonymous Mutations)	121
5.6	Performance Evaluation of the Classifiers (Synonymous mutations)	122
5.7	IDM Features and their Description	127
5.8	Predictive Performance of the Classifiers (Insertion, deletion and duplication Mutations)	131
5.9	Performance Evaluation of the Classifiers (IDD mutations)	132
5.10	Feature Values and their Description	136
5.11	Predictive Performance of the Classifiers (Splicing Mutations)	139
5.12	Performance Evaluation of the Classifiers (Splicing mutations)	140
5.13	Predictive Performance of the Classifiers (Pooled Features)	146
5.14	Predictive Performance of the Classifiers (After feature selection)	148
5.15	Performance evaluation of the Classifiers (After Feature selection)	148

5.16	Comparative study of Prediction Accuracy Before and After Feature Selection	149
6.1	Predictive Performance of the LibD3C Classifier (Non Synonymous mutation)	154
6.2	Predictive Performance of the LibD3C Classifier (Synonymous mutation)	155
6.3	Predictive Performance of the LibD3C Classifier (Insertion, deletion and duplication Mutations)	157
6.4	Predictive Performance of the LibD3C Classifier (Splicing mutation)	158
6.5	Predictive Performance of the LibD3C Classifier (AGM Dataset)	160
6.6	Comparison of LibD3C Classifiers	161
6.7	Predictive Accuracy Comparison with Supervised Learning Algorithms	162
7.1	Predictive Performance of the Classifiers with Nucleotide Mapping	169
7.2	Codon and its Integer Identifiers	171
7.3	Predictive Performance of the DNN with Codon Mapping	175
7.4	Predictive Performance of the ANN Classifier	177
7.5	Performance Comparison of Shallow Net with Deep Net (with hidden layers ranging from 3 to 8 in Nucleotide Mapping Scheme)	178
7.6	Performance Comparison of Shallow Net with Deep Net (with hidden layers ranging from 3 to 8 in Codon Mapping Scheme)	179