

TABLE OF CONTENTS

DISSERTATION	i
ENDORSEMENT PAGE	iv
DISSERTATION	iv
DECLARATION	v
DEDICATION	vi
PREFACE	viii
TABLE OF CONTENTS	ix
LIST OF FIGURES	xii
LIST OF TABLES	xiii
ATTACHMENT	xiv
GLOSSARY	xv
ABBREVIATION & ACRONYMS	xvi
ABSTRACT	xvii
CHAPTER I INTRODUCTION	1
1.1 Problem and Background	1
1.2 Problem Statements	7
1.3 Research Questions	8
1.4 Scope and Limitation of Research	9
1.5 Research Objective	9
1.6 Research Contribution	9
CHAPTER II LITERATURE REVIEW	11
2.1 Leukemia Cancer	11
2.2 Machine Learning Approach	13
2.2.1 Support Vector Machine	14
2.2.2 Rule Based Machine Learning	14
2.2.3 Data Mining	15
2.2.4 Deep Learning	16
2.2.5 Practical Swarm Optimization & Principal Component Analysis	17
2.2.6 Decision Tree & Random Forest	18
2.3 Voting Classifier	19
2.4 Microarray and Gene Expression	21
2.5 Ensemble Learning	25
CHAPTER III THEORETICAL FUNDAMENTALS	31
3.1 Theoretical Background	31
3.2 Definition Ensembles Classifier	33
3.3 Machine Learning Methods	36
3.3.1 Support Vector Machine Classifier	36
3.3.2 Gradient Boosting Machine	37
3.3.3 Decision Tree Classifier	37
3.3.4 Naive Bayes Classifier	37

3.3.5	Random Forest Classifier	38
3.3.6	Logistic Regression Algorithm.....	38
3.4	Different Proposed Approaches	38
3.4.1	A Fast and Lightweight AutoML Library (FLAML)	38
3.4.2	Voting Classifier.....	39
3.4.3	Synthetic minority oversampling technique (SMOTE).....	41
3.4.4	Principal component analysis (PCA).....	41
3.5	Extended Voting Classifier and Machine Learning architecture.....	42
3.6	Types of Machine Learning Architecture.....	42
3.6.1	Supervised Learning	43
3.6.2	Unsupervised Learning	44
3.6.3	Reinforcement Learning	44
3.7	Machine Learning Architecture.....	44
3.7.1	Acquiring Data.....	44
3.7.2	Processing of Data	45
3.7.3	Modeling of Data.....	45
3.7.4	Execution	45
3.7.5	Deployment.....	46
3.8	Model Logistic Regression Algorithm	46
3.9	DNA Microarray	47
3.10	Microarray technology	47
3.11	Types of Leukemia.....	48
3.11.1	Acute Myeloid Leukemia (AML).....	49
3.11.2	Acute Lymphocytic Leukemia (ALL)	49
3.11.3	Chronic Myeloid Leukemia (CML).....	49
3.11.4	Chronic Lymphocytic Leukemia (CLL).....	49
CHAPTER IV RESEARCH METHODS & IMPLEMENTATION OF EXPERIMENTS EVALUATION.....		50
4.1	Materials and methods.....	50
4.2	Overview of the proposed approach.....	50
4.3	Datasets for Leukemia Cancer.....	51
4.4	Description of the datasets.....	52
4.5	Classification	63
4.6	Supervised machine learning models	63
4.7	Proposed voting classifier strategy.....	63
4.8	Algorithms for Leukemia Classification	64
4.8.1	Decision Tree (DT).....	65
4.8.2	Naïve Bayes (NB).....	66
4.8.3	Random Tree (RT).....	66
4.8.4	Linear Regression (LR)	67
4.8.5	Support Vector Machine (SVM)	67
4.8.6	Gradient Boosting Machines (GBMs).....	68
4.8.7	Hybrid voting classifier model	68
4.9	Ensemble with Hard and Soft voting.....	69
4.10	Research Stages	69
4.10.1	Background Review.....	70
4.10.2	Problem Formulation & Design.....	70
4.10.3	Implementing & Evaluation	71
4.11	Training of Dataset.....	71
4.12	Evaluation of the Model Extended Voting Classifier.....	72
4.13	Cross-validation.....	72



4.14 Methodology	73
4.15 Preprocessing of the datasets.....	74
4.16 Combination of SMOTE and PCA for balancing.....	75
4.17 Proposed architecture of LDSVM.....	75
4.18 Evaluation Criteria	76
4.18.1 Criteria for Accuracy	77
4.18.2 Criteria for Precision.....	77
4.18.3 Criteria for Recall	77
4.18.4 Criteria for F1-Score.....	78
CHAPTER V ANALYSIS AND RESULTS.....	79
5.1 Result and discussions.....	79
5.2 Models' performance on original leukemia dataset	79
5.3 Significance of proposed approach	83
5.3.1 Models' performance results on leukemia_GSE71935 dataset	83
5.3.2 Models' performance on leukemia_GSE28497 dataset	85
5.3.3 Models' performance on Leukemia_GSE63270 dataset	86
5.3.4 Models' performance on the Leukemia_GSE71449 dataset	87
5.4 Models' performance comparison results on leukemia datasets	88
5.5 Performance analysis of the current proposed approach.	90
CHAPTER VI CONCLUSION & RECOMMENDATION	106
6.1 Conclusion.....	106
6.2 Recommendations	108
REFERENCES	109
ATTACHMENT	123

LIST OF FIGURES

Figure 1.1.	Different applications procedures of ML and DL for multiple leukemia research.....	4
Figure 3.1.	Supervised learning process	34
Figure 3.2.	Voting classifier	39
Figure 3.3.	Main computational techniques for ML	43
Figure 3.4.	Architecture for machine learning systems	44
Figure 3.5.	DNA microarray	47
Figure 3.6.	Types of leukemia disease.....	48
Figure 4.1.	Classification approach for leukemia cancer based on the proposed approach	51
Figure 4.2.	Five different types of leukemia presentation on Leukemia_GSE9476 genes dataset.....	52
Figure 4.3.	Based on classes variations of different gene data presentations	52
Figure 4.4.	Sample shot of Leukemia_GSE9476 dataset	55
Figure 4.5.	Sample shot of Leukemia_GSE71935 dataset	56
Figure 4.6.	Sample shot of Leukemia_GSE28497 dataset	58
Figure 4.7.	Sample shot of Leukemia_GSE63270 dataset	60
Figure 4.8.	Sample shot of Leukemia_GSE71449 dataset	61
Figure 4.9.	Research stages.....	70
Figure 4.10.	The methodology adopted for classification using ML.....	71
Figure 4.11.	Voting classifier approach.....	72
Figure 4.12.	Classification of leukemia cancer based on ML models and ensemble LDSVM (LR+DT+SVM).....	73
Figure 4.13.	Proposed approach structure of ensemble LDSVM model	76
Figure 5.1.	Confusion matrix comparative analyses of machine learning model.....	80
Figure 5.2.	Result of different classifier based on parameters performance	82
Figure 5.3.	Number of components against the cumulative variance.....	83
Figure 5.4.	Confusion matrix of LDSVM on Leukemia_GSE71935 dataset.....	84
Figure 5.5.	Confusion matrix of LDSVM on Leukemia_GSE28497 dataset.....	86
Figure 5.6.	Confusion matrix of LDSVM on Leukemia_GSE63270 dataset.....	87
Figure 5.7.	Confusion matrix of LDSVM on Leukemia_GSE71449 dataset.....	88
Figure 5.8.	Classification accuracy	90

LIST OF TABLES

Table 2.1.	Various techniques of the voting classifier with machine learning and deep learning approach	28
Table 4.1.	Detailed dataset information Leukemia_GSE9476	53
Table 4.2.	Detailed dataset information Leukemia_GSE71935	55
Table 4.3.	Detailed dataset information Leukemia_GSE28497	57
Table 4.4.	Detailed dataset information Leukemia_GSE63270	59
Table 4.5.	Detailed dataset information Leukemia_GSE71449	60
Table 4.6.	Online web access links of datasets	61
Table 4.7.	Range and setting of models hyperparameters used for tuning.....	64
Table 5.1.	Comparative result of proposed ensemble LDSVM and machine learning models	81
Table 5.2.	Comparative result of proposed ensemble LDSVM and machine learning models (Balancing).....	82
Table 5.3.	Comparative result of proposed ensemble LDSVM and machine learning models on Leukemia_GSE71935 dataset.....	84
Table 5.4.	Comparative result of proposed ensemble LDSVM and machine learning models on Leukemia_GSE28497 dataset.....	85
Table 5.5.	Comparative result of proposed ensemble LDSVM and machine learning models on Leukemia_GSE63270 dataset.....	86
Table 5.6.	Comparative result of proposed ensemble LDSVM and machine learning models on Leukemia_GSE71449 dataset.....	87
Table 5.7.	Comparative overall accuracy results of machine learning models on various datasets.....	89
Table 5.8.	Comparison results from the literature of the proposed LDSVM classifier ensemble approach with present approaches based on datasets	91
Table 5.9.	The accurate model with different parameters on Leukemia_GSE9476 ...	97
Table 5.10.	Accurate model with different parameters on Leukemia_GSE63270.....	98
Table 5.11.	Accurate model with different parameters on Leukemia_GSE71935.....	100
Table 5.12.	Accurate model with different parameters on Leukemia_GSE28497.....	102
Table 5.13.	Accurate model with different parameters on Leukemia_GSE71449.....	104

ATTACHMENT

Attachment	Description
Attachment A	List of Publication
Attachment B	Coding
Attachment C	Coding
Attachment D	Dataset Tables

GLOSSARY

No.	Terms	Description
1.	Microarray	Microarray technology is a rapidly evolving technique for studying the expression of a large number of genes simultaneously.
2.	DNA	Deoxyribonucleic acid (DNA) is a vast molecule containing the genetic code unique to each individual because it provides the instructions for creating all of the proteins in our body.
3.	Genes	The fundamental unit of heredity occupies a specific chromosomal position.
4.	Cancer	Cancer is a disease that requires medical attention. Some of the body's cells develop in an unregulated manner, and they have the potential to invade other parts of the body.
5.	Leukemia	Leukemia is a type of cancer, and Bone marrow is one of the tissues affected by leukemia, which is a malignancy of the blood-forming tissues.
6.	Bone Marrow	The marrow of the bone is a spongy material that is found in the interior of the bones. It is responsible for the formation of bone.
7.	ALL	Among children under five, acute lymphocytic leukemia (ALL) is the most frequent leukemia.
8.	AML	Acute myelogenous leukemia (AML) is a kind of leukemia that affects many people. It can affect both children and adults.
9.	CLL	In adults, chronic lymphocytic leukemia (CLL) is the most common kind of leukemia
10.	CML	Chronic myelogenous leukemia is the kind of leukemia in adults who are more likely than children to develop chronic myelogenous leukemia.
11.	Bone_Marrow_CD34	It is the dataset group type.
12.	PBSC_CD34	It is the dataset group type.

ABBREVIATION & ACRONYMS

No.	Terms	Description
1.	LDSVM	Logistic Regression, Decision Tree, Support Vector Machine
2.	WBCs	“White blood cells”
	RBC	“Red Blood Cell”
3.	PB	“Peripheral Blood”
4.	ALL	“Acute Lymphoblastic Leukemia”
5.	AML	“Acute Myeloid Leukemia”
6.	CLL	“Chronic Lymphocytic Leukemia”
7.	BM	“Bone Marrow”
8.	MRI	“Magnetic Resonance Imaging”
9.	SMOTE	“Synthetic Minority Oversampling Technique”
10.	PCA	“Principal component analysis”
11.	ML	“Machine Learning”
12.	FLAML	“A Fast and Lightweight AutoML Library “