



FINAL YEAR PROJECT REPORT

PEDIATRIC CHEST DISEASE PREDICTION

**In fulfillment of the requirement
For degree of
BS (COMPUTER SCIENCES)**

By

**RAAHIMA IRFAN
MUSADDIQ ALI SADIQ**

**70110 (BSCS)
70145 (BSCS)**

SUPERVISED

BY

MS.TOOBA MEHTAB

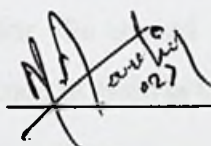
BAHRIA UNIVERSITY (KARACHI CAMPUS)

SPRING-2024

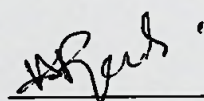
DECLARATION

We hereby declare that this project report is based on our original work except for citations and quotations which have been duly acknowledged. We also declare that it has not been previously and concurrently submitted for any other degree or award at Bahria University or other institutions.

Signature :

Name : MUSADDIQ ALI SADIQReg No. : 70145

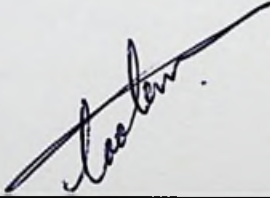
Signature :

Name : RAAHIMA IRFANReg No. : 70110Date : 14/06/2024

APPROVAL FOR SUBMISSION

We certify that this project report entitled “**PEDIATRIC CHEST DISEASE PREDICTION**” was prepared by **MUSADDIQ ALI SADIQ** and **RAAHIMA IRFAN** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Science in Bachelors’ of Science in Computer Science at Bahria University, Karachi.

Approved by,

Signature : _____

Supervisor: Ms. TOOBA MEHTAB

Date : 14/06/24

ACKNOWLEDGEMENTS

We would like to thank everyone who had contributed to the successful completion of this project. We would like to express my gratitude to my project supervisor, Mr. TOORAJ MEHTA for his guidance, advice, patience and his constant support throughout the development of this report.

In addition, we would like to express our gratitude to our loving parents

The copyright of this report belongs to Bahria University according to the Intellectual Property Policy of Bahria University BUORIC-P15 amended on April 2019. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this report.

ACKNOWLEDGEMENTS

We would like to thank everyone who had contributed to the successful completion of this project. We would like to express my gratitude to my project supervisor, Ms. TOOBA MEHTAB for her invaluable advice, guidance and her enormous patience throughout the development of the research.

In addition, we would also like to express our gratitude to our loving parents and friends who have helped and given us encouragement.

PEDIATRIC CHEST DISEASE PREDICTION

ABSTRACT

This study addresses the classification of pediatric chest X-ray images into three diagnostic categories: Infiltration, Effusion, and Atelectasis. Utilizing the NIH Chest X-ray dataset filtered for patients aged 12 or younger, we employ convolutional neural networks (CNNs) for feature extraction and rigorously compare various machine learning and deep learning models, including Random Forest, Logistic Regression, SVM, Decision Tree, and DenseNet121. A key contribution of our research is the use of a pre-trained VGG16 model for feature extraction, followed by the application of a Random Forest classifier, which achieved a high classification accuracy. Our experimental methodology involves addressing class imbalance through dataset balancing techniques and evaluating models on differently sized datasets. Results indicate that larger datasets substantially enhance model performance, with DenseNet121 attaining the highest overall accuracy and AUC scores. Notably, our VGG16-augmented Random Forest model also demonstrated robust performance, highlighting the effectiveness of our feature extraction and classification pipeline. The findings of this study underscore the necessity of adequate training data and validate the superiority of deep learning models in handling complex medical imaging tasks. This research underscores the potential of advanced machine learning techniques in improving pediatric chest X-ray classification, thus enhancing diagnostic accuracy and supporting clinical decision-making. Future research directions will explore data augmentation, hyperparameter optimization, ensemble methods, and strategies to improve model interpretability and generalizability.

TABLE OF CONTENTS

DECLARATION	ii
APPROVAL FOR SUBMISSION	iii
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF SYMBOLS / ABBREVIATIONS	xiii
LIST OF APPENDICES	xiv

CHAPTER

1	INTRODUCTION	1
	1.1 Background	1
	1.2 Problem Statement	1
	1.3 Aims and Objectives	1
	1.4 Scope of Project	2
2	LITERATURE REVIEW	3
	2.1 Theoretical Background	3
	2.2 Machine learning and Deep Learning Approach	3
	2.2.1 Convolutional Neural Networks	3
	2.2.2 Ensemble Methods	4
	2.3 Related Work	5
	2.4 Research Gap	5
	2.5 Dataset characteristics	6
	2.5.1 Comparative Analysis	6
3	METHODOLOGY	8
	3.1 Proposed Methodology	8

	3.1.1	Data Acquisition	8
	3.1.2	Data Cleaning	9
	3.1.3	Class Balance	10
	3.1.4	Feature Extraction	10
	3.1.5	Feature Selection using PCA	10
	3.1.6	Random Forest Classifier	11
4		EXPERIMENTAL SETUP	12
	4.1	Overview	12
	4.2	Programming Environment	12
	4.2.1	Platform	12
	4.2.2	Programming Language	12
	4.2.3	Hardware	12
	4.3	Libraries and Tools	12
	4.3.1	Data Manipulation	12
	4.3.2	Visualization	13
	4.3.3	TensorFlow and Keras	13
	4.3.4	Scikit-learn	13
	4.3.5	Kaggle Datasets	13
	4.4	Data Preparation	13
	4.4.1	Dataset Description	13
	4.4.2	Data Loading and Initial Processing	13
	4.4.3	Class Distribution and Balancing	14
	4.5	Feature Extraction	14
	4.6	Dimensionality reduction	14
	4.7	Data Splitting and Model Training	14
	4.8	Hyperparameter Tuning	16
	4.9	Evaluation Matrices	16
	4.10	Results	16
5		RESULTS	17
	5.1	Overview	17
	5.2	Results with Equal Large dataset	17

5.2.1	Random Forest	17
5.2.2	Comparison with Machine Learning Algorithm	20
5.2.3	Comparison with Deep Learning	21
5.3	Results with Equal but Less Images	21
5.3.1	Random Forest	21
5.3.2	Comparison with Machine Learning Algorithms	23
5.3.3	Comparison with Deep Learning Algorithms	23
6	CONCLUSION	24
6.1	Overview	24
6.2	Summary of Findings	24
6.2.1	Dataset Size and Model Performance	24
6.2.2	Class Imbalance	25
6.2.3	Machine Learning vs. Deep Learning	25
6.3	Limitations	25
6.3.1	Dataset Imbalance and Bias	25
6.3.2	Computational Resources	25
6.3.3	Generalizability	25
6.3.4	Concluding Remarks	26
7	FUTURE WORK	27
7.1	Introduction	27
7.2	Advanced Data Augmentation	27
7.2.1	Generative Adversarial Networks	27
7.2.2	Data Augmentation Techniques	27
7.3	Hyperparameter Optimization	27
7.3.1	Bayesian Optimization	27
7.3.2	Genetic Algorithms	28
7.4	Ensemble Learning	28
7.4.1	Combining Models	28
7.4.2	Stacking and Blending	28
7.5	Explain ability and Interpretability	28
7.5.1	Visual Explanations	28

	7.5.2	Feature Importance	28
7.6		Cross-Dataset Validation	29
	7.6.1	Diverse Datasets	29
	7.6.2	Real-World Clinical Data	29
7.7		Integration with Clinical Workflows	29
	7.7.1	Deployment in Clinical Settings	29
	7.7.2	Continuous Learning Systems	29
7.8		Collaborative Research and Open Science	29
	7.8.1	Collaborative Efforts	29
	7.8.2	Open-Source Contributions	30
7.9		Concluding Remarks	30
REFERENCES			31
APPENDICES			36