ENHANCING RANDOM FOREST MODEL FOR CANCER DETECTION: A MULTI-OBJECTIVE HYPERPARAMETER TUNING STRATEGY

BY

Md Saykot Islam
ID: 201-15-13993

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled **"ENHANCING RANDOM FOREST MODEL FOR CANCER DETECTION: A MULTI-OBJECTIVE HYPERPARAMETER TUNING STRATEGY"**, submitted by **Md Saykot Islam** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 1/25/2024.

BOARD OF EXAMINERS

Chairman

Dr. Md. Zahid Hasan (ZH)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

25.01.2024

Dr. Fizar Ahmed (FZA)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Abdus Sattar (AS)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammed Nasir Uddin (DNU)
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Md. Ferdouse Ahmed Foysal, Lecturer Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Saykot

Md Saykot Islam
ID: 201-15-13993
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, I would like to thank Almighty Allah from the bottom of my heart for His divine blessing, which allowed us to successfully complete the final year thesis.

To **Md. Ferdouse Ahmed Foysal, Lecturer**, Department of CSE, Daffodil International University, Dhaka, I sincerely thank you and wish to express my profound gratitude. To complete this project, my supervisor must have deep knowledge of and a strong interest in the field of "machine learning." The completion of this project has been made possible by his endless patience, scholarly guidance, ongoing encouragement, constant and energetic supervision, constructive criticism, insightful advice, reading numerous subpar drafts, and correcting them at every stage.

I would like to extend my sincere gratitude to Professor **Dr. Sheak Rashed Haider Noori**, **Head (In-Charge) of the CSE Department** at Daffodil International University, as well as other faculty members and staff for their valuable help in completing my project.

I want to express my gratitude to every student at Daffodil International University who participated in this discussion while finishing their coursework.

Lastly, I must respectfully thank our parents for their ongoing assistance and patience.

ABSTRACT

Early and accurate cancer detection is crucial for improving patient outcomes from this lethal disease, but traditional methods often lack sensitivity or specificity. Machine learning algorithms, particularly Random Forests, offer promising tools for analyzing medical data and achieving this goal. However, the performance of Random Forests is heavily dependent on the appropriate configuration of hyperparameters, requiring an optimal configuration for accurate prediction. This study proposes a novel multi-objective hyperparameter tuning strategy to enhance the effectiveness of Random Forests for cancer detection. The research starts off with single-goal hyperparameter tuning targeted on accuracy, observed by an evaluation of performance the usage of five-fold move-validation. The results show a cross-validated training accuracy of 0.96 and a test accuracy of 0.97, effectively addressing the issue of overfitting. Subsequently, the methodology advances a state-of-the-art optimization algorithm, inclusive of a multi-objective algorithm or a particle swarm optimization, to explore the hyperparameter area efficiently. The proposed strategy aims to construct a Random Forest model that not only delivers accuracy but also maintains equilibrium across diverse performance aspects in cancer detection. To reap this, a multi-goal optimization algorithm is integrated with the hyperparameter tuning technique, enabling the exploration of numerous solutions across the Pareto front. This approach enhances the version's potential to parent diffused patterns indicative of cancerous conditions whilst minimizing false positives and fake negatives. The proposed multi-objective hyperparameter tuning approach for Random Forest models is a significant development in the field of cancer detection. In the end, it uses machine learning to improve healthcare outcomes by paving the way for more precise, understandable, and clinically relevant cancer diagnoses. It also highlights the significance of thorough hyperparameter optimization techniques in boosting model efficacy.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER **PAGE**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Breast cancer continues to be the second leading cause of cancer-related death among women, despite an escalating incidence among women. Each year, the World Health Organization (WHO) estimates 1.38 million new cases of breast cancer and 458,000 deaths from it. This malignancy is a common disease across both developed and developing nations, with survival rates varying significantly worldwide. As compared with North America, Sweden, and Japan, middle-income countries experience survival rates of 60% or less, and low-income countries experience survival rates of less than 40%. The integration of expert systems and machine learning techniques in medical diagnosis has witnessed a gradual but impactful rise. While the evaluation of patient data and expert decisions remains paramount in diagnosis, artificial intelligence systems offer valuable assistance to medical experts. Automatic diagnostic systems play a crucial role in mitigating potential errors in diagnosis, enabling a more expedited and detailed examination of medical data. The purpose of this study is to develop a computer-aided diagnostic system that can differentiate benign from malignant breast tumors. The proposed method employs a two-stage approach, combining a backward elimination feature selection method with a learning algorithm, specifically the random forest model. The first stage involves a data reduction process that prepares the dataset for the subsequent random forest algorithm, thereby optimizing prediction time. The selected feature set not only enhances the system's interpretability but also contributes to a more concise rationale for diagnostic outcomes. The dataset utilized for this study is the Wisconsin Diagnosis Breast Cancer Orginal Dataset obtained from the University of California at Irvine (UCI) Machine Learning Repository. Widely recognized in the research community, this dataset allows for performance comparisons with other studies focusing on expert systems and machine learning methods for breast cancer diagnosis.

1.2 Motivation

This study is inspired by the pressing need to address the rising incidence of breast cancer worldwide, which is the leading cause of cancer-related deaths among women. This urgency is highlighted by World Health Organization (WHO) statistics, which indicate an increasing number of new cases and yearly deaths. Although overfitting is still a problem, machine learning more especially, the Random Forest model is being emphasized for the detection of breast cancer because of its potential. The main goal is to develop a powerful computer-aided diagnosis system that can accurately identify breast tumors that are benign or malignant. A complex hyperparameter tuning strategy is required to overcome overfitting complexities in order to accomplish this goal.The potential effect is significant, offering improved precision and promptness in breast cancer diagnoses, thereby positively affecting patient results. Moreover, this research aims to make a valuable contribution to the worldwide effort to combat breast cancer through the creation of a versatile diagnostic tool. The objective is to develop a system that surpasses differences in survival rates worldwide, offering a dependable and flexible solution applicable to various datasets and situations. The primary objective of this study is to contribute to the improvement of medical diagnostics and make significant advancements in the ongoing fight against breast cancer, in line with the overarching goal of enhancing public health.

1.3 Rationale of the Study

The motivation behind this research is to improve the effectiveness of cancer detection through the utilization of the Random Forest model, with a specific focus on breast cancer, which is a substantial worldwide health issue. Acknowledging the necessity for enhanced diagnostic instruments, the justification is founded on the conviction that by optimizing the Random Forest algorithm via a multi-objective hyperparameter tuning methodology, a more equitable and stable model can be generated to ensure precise cancer detection.

Despite challenges observed with default hyperparameters, the study integrates an effective hyperparameter tuning strategy. By incorporating the Multi-Objective Hyperparameter Optimization (MOHPO) approach, the emphasis on accuracy, TP rate, FP rate, and the

AUC curve is raised. This approach aims to improve the RF model's performance metrics, anticipating outcomes that could dramatically contribute to the development of machine learning applications in medical diagnostics.

1.4 Research question

Developing a precise, succinct, and targeted research question is an essential first step in any investigation. It gives the field of study a clear outline and a clear sense of direction. In order to formulate a workable, efficient, and precise plan of action for resolving the research problem, investigators ought to formulate follow-up inquiries that clarify their ideas and conclusions. This iterative process helps to improve the research, directing the study toward significant discoveries and useful conclusions.

- Is it possible to use an optimized Random Forest to detect cancer?
- Can the field of cancer detection be improved by this work?
- How can a multi-objective hyperparameter tuning strategy be effectively employed to enhance the accuracy, specificity, and sensitivity of Random Forest models for improved cancer detection, ultimately making diagnoses more precise and clinically relevant?

1.5 Expected Outcome

I have discussed my research proposal, which is based on the research question, in this section. Our goal is to create a well-known model for the suggestion. The following is what the researchers hope to achieve with their study.

- Optimized Random Forest Implementation: It is anticipated that an optimized Random Forest model will be implemented successfully to detect cancer, outperforming default configurations.
- Better Cancer Detection: By using the optimized Random Forest model, it is anticipated that cancer detection will be more accurate and efficient.

1.6 Research Management

I methodically implemented the intended architecture with efficient project management to wrap up the research project. The following is a summary of the project flow:

I. Logistics Searches: Started by gathering relevant research of relevant field.

II. Study Selection: Thoroughly selected studies aligned with our themes.

III. Analytical Research: Performed comprehensive analyses to comprehend detection systems.

IV. Data Collection: Obtained particular data from the Breast Cancer Wisconsin (Original) [1].

V. Dataset Preprocessing: Preprocessed data with classification algorithm.

VI. Implementing the Classifier: Applied classifier

VII. Training and Testing: The classifier was thoroughly tested for accuracy in real-world scenarios.

VIII. Documentation: Thoroughly recorded research results and discoveries.

1.7 Lay Out of the Report

The study's motivation and reasoning are presented in Chapter 1, which also acts as an introduction. It includes the study questions, expected outcomes, project management, financial information, and a summary of the project's organizational structure.

Chapter 2 discusses the research background, related studies, challenges, comparative analysis, and preliminary work are covered. It delineates the extent of the recognized issue.

Chapter 3, a theoretical analysis is carried out, describing the data collection procedure, the algorithmic approach to machine learning, and the project progress. It also covers the requirements for implementation.

Chapter 4 the experimental results are presented, the findings are discussed, and the project performance is evaluated.

Chapter 5 describes the research's societal impact, environmental implications, and ethical dilemmas are examined.

Chapter 6 the study is concluded, which also highlights its limitations and offers suggestions for further research.

CHAPTER 2
BACKGROUND

2.1 Preliminaries and Terminologies

This section explores fundamental concepts that are necessary for a thorough comprehension of the background of our cancer detection research. To maintain a consistent and accurate language throughout our investigation, I started by defining important terms that are essential to the field. The purpose of these definitions is to provide clarification and avoid confusion while navigating the intricacies of cancer detection techniques. Beyond terminology, let look into preliminary details that establish the framework for our research. This comprises background data particular to cancer detection, theoretical frameworks, and basic concepts. By carefully analyzing these preliminary findings, established the groundwork for a more thorough investigation of our research questions and goals concerning cancer detection.

2.2 Related Works

Previously several research works have been done on cancer detection by implementing Random Forest. In machine learning, choosing the best classifier is essential, especially when dealing with medical datasets. Data scientists have used a variety of algorithms, with excellent outcomes.

- The authors of the paper by Kaur et al. [2] use Bayesian Optimization to optimize the model's hyperparameters in order to maximize the AUC score. Nevertheless, because their suggested method is single-objective, it might not adequately address the multi-objective nature of the current issue. As a result, there are few studies that apply Multi-Objective Hyperparameter Optimization (MOHPO) methods to the diagnosis of breast cancer. In this paper , with respect to the other models that were taken into consideration, the BSense model performed better, achieving 83.9%, 87.3%, 91.1%, and 80.1% Area Under Curve (AUC) for the TCGA, METABRIC, Metabolomics, and RNA-seq datasets, respectively.

- Delen et al. Lu [3] used a dataset of 202,932 patient records with breast cancer patients divided into two groups: malignant (93,273) and benign (109,659). With the help of the Random Forest algorithm, the predictive model showed an astounding 93% accuracy rate. This demonstrates how well Random Forest performs as a strong classifier for precisely predicting the outcomes of breast cancer based on the given dataset, highlighting its potential in medical applications.
- Awotunde et al. [4] explore the laborious process of hyperparameter tuning for a blood cancer classifier in their most recent work. They are primarily interested in optimizing the suggested model's accuracy. This project emphasizes how important it is to optimize the model's parameters in order to attain improved performance, especially when it comes to blood cancer detection. The thorough methodology used by Awotunde and colleagues advances the continuous endeavors to improve the accuracy and efficiency of classifiers in the field of medical diagnostics.
- Using the WBCD dataset, Mamta Jadhav and Zeel Thakkar [5] implemented the Random Forest algorithm. They obtained a notable accuracy rate of 95% by dividing the dataset into a training set (75%) and a testing set (25%). This demonstrates the Random Forest algorithm's strong performance during their investigation, especially in correctly differentiating between benign and malignant cases in the WBCD dataset.
- Cameron Wolfe [6]used the sklearn breast cancer dataset, which included 569 patients, to demonstrate a Random Forest method. Notably, he used 50 decision trees and eliminated variables that showed high correlation (>0.9). The result showed that the Random Forest algorithm was effective in producing accurate results for breast cancer diagnosis based on the particular dataset, with an impressive accuracy result of 96.71%.
- Naji et al. [7]provide data in the literature that highlight the effectiveness of the Support Vector Machine and Random Forest algorithms in routinely obtaining accuracy levels above 96% when it comes to the identification of malignant tumors. This data shows how well these machine-learning techniques work in the field of tumor detection.

- Azar and El-Metwally [8] contribute to the literature in the field of breast cancer diagnosis by demonstrating accuracies exceeding 95% using various Decision Tree classifiers.
- Desai and Shah [9] supplement this with a Multilayer Perceptron (MLP) classifier, achieving an accuracy of 91.9% on the Wisconsin breast cancer dataset. This literature emphasizes the versatility of different classifiers in accurately diagnosing breast cancer, emphasizing the importance of tailored approaches based on different datasets.
- A detailed analysis of the accuracy values corresponding to four well-known machine learning classification models (LR, KNN, Random Forest Tree - RDT, and SVM) was carried out by Chaurasiya et al. [10]. The evaluation centered on how well they performed using the WDBC dataset. Interestingly, Random Forest Tree (RDT) achieved the highest accuracy of all the classifiers that were evaluated, with an astounding 95%. This demonstrates Random Forest Tree's remarkable performance when compared to other well-known classification models in the context of the WDBC dataset.
- The Random Forest algorithm was proposed by Sarinder Kaur Dhillon and Pietro Lio [4] for use in cancer diagnosis prediction. They used an 8,066-record dataset that they obtained from the University Malaya Medical Centre for their study. Using the Random Forest algorithm, dataset was split into atraining set (70%, 5,646 records) and a testing set (30%, 2,420 records) in order to evaluate the model's performance. The model's obtained accuracy of 82.2% highlights the effectiveness of Random Forest in predicting cancer diagnoses using data from the University Malaya Medical Centre.

2.3 Scope of the problem

This study aims to investigate the particular parameters and dimensions involved in improving the Random Forest model for cancer detection through a multi-objective hyperparameter tuning approach. It provides a precise definition of the problem's scope by outlining the parameters, factors, and considerations that are included in the scope of this investigation. The research framework becomes more precise by defining the scope, which allows for a targeted and methodical examination of the Random Forest model's optimization for cancer detection while taking multi-objective hyperparameter tuning into account. and determine the trade-offs between various objectives, which are crucial for making medical diagnoses.

2.4 Challenges

The principal obstacle that I faced during our investigation was gathering datasets and choosing a suitable algorithm, which we finally determined to be the traditional Random Forest. A second problem surfaced concerning efficient Random Forest algorithm optimization. I employed hyperparameter tuning, a procedure that added complications of its own, to address this. At the same time, there was a big challenge in figuring out relevant and essential features for our prediction model. By using a methodical approach that included thorough cleaning and pre-processing of the raw dataset, I was able to overcome this challenge and use only the features that were necessary for our prediction model. Moreover, the inclusion of cross-validation introduced an extra degree of complexity. It took a methodical and well-considered approach to strike a balance between the complexity of feature selection and hyperparameter optimization and the requirement for a robust model evaluation. The overall effectiveness of our research objectives was aided by this thorough methodology, which sought to improve the Random Forest algorithm's performance.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this section, I briefly describe the process of our suggested model for cancer detection classification, with a focus on using the Random Forest algorithm. Equations, graphs, and thorough explanations are used to explain important elements such as data gathering, processing, and the suggested model. Our cancer detection dataset is applied to the Random Forest algorithm and is carefully tuned through Hyperparameter Tuning to achieve optimal performance. This chapter's last section explores our project's statistical underpinnings and clarifies its conceptual structure. In addition, I clarify the particular implementation needs that are essential to the proper operation of our suggested Random Forest-based cancer detection model. This thorough explanation guarantees a clear comprehension of the methodological stages involved in creating and applying our model, emphasizing the critical function of the Random Forest algorithm in obtaining reliable and accurate cancer classification results.

3.2 Research Subject and Instrumentation

Topic for research includes those that help us easily to understand topic concepts. Design model execution, processing dataset, dataset collection, training model, and altering it in response to how it performs. In Another section, tools, is primarily concerned with techniques and the methods employed. I programmed on the Windows platform for this proposed project. Pandas, NumPy, Skit Learn, Seaborn, Matplotlib, and other Python libraries. Throughout the training and testing process, Jupyter Notebook is used.

3.3 Workflow

In this section will describes the working flow that I used to complete my project. InstrumentData gathering, processing, model selection, and assessment of the outcome and the challenge's future potential were all accomplished with success.ntation

Stage 1 - Data Collection: Wisconsin Breast Cancer Orginal

The primary dataset for cancer detection is an original dataset taken from the UCI Machine Learning Repository [1]. The chosen dataset is thoroughly tested for reliability and relevance to the research objectives. A thorough examination of the dataset's characteristics, including features and target variables, is carried out. It has 699 instances, 9 features, and 1 target class which is benign or malignant. Where the number of benign is 458 and the number of malignant is 241.

Stage 2 - Data Processing: The Breast Cancer Wisconsin Original dataset [1] is precisely processed at this stage to ensure its suitability for the training and evaluation of the Random Forest model for breast cancer detection.

- **Feature Selection:** To select the most relevant features, here used the Random Forest model's inherent feature importance scores. And worked with the top 5 important features.
- **Missing Values Handling:** Missing values are handled using simple imputer using mean strategy.
- **Splitting Datasets:** Splited the datasets into 25% test sets and 75% train set.

Stage 3 - Model Selection: From various ML models I choosed the Random Forest and Hyperparameter Optimization (HPO) algorithms. I used Grid Search to optimize the output of Random Forest algorithms.

Stage 4 - Performance Evaluation: During this critical phase of performance evaluation, a comprehensive analysis is conducted using a variety of important metrics and analytical tools on the improved Random Forest model intended for cancer detection. Measures like recall, accuracy, True positive rate, false positive rate, and F1 score are used to give a thorough picture of the model's performance. Furthermore, the discriminating power of the model is quantified by the Receiver Operating Characteristic (ROC) curve, which shows the trade-off between different objectives when the Area Under the ROC Curve is calculated. It shows the comparison between default parameters, single objective hyperparameters, and multiobjective hyperparameters used for tradeoffs and gaining good accuracy through 5-fold cross-validation.

Stage 5 - Conclusion and Future work: This section will outline the project's future scope and provide a brief rundown of the entire procedure.

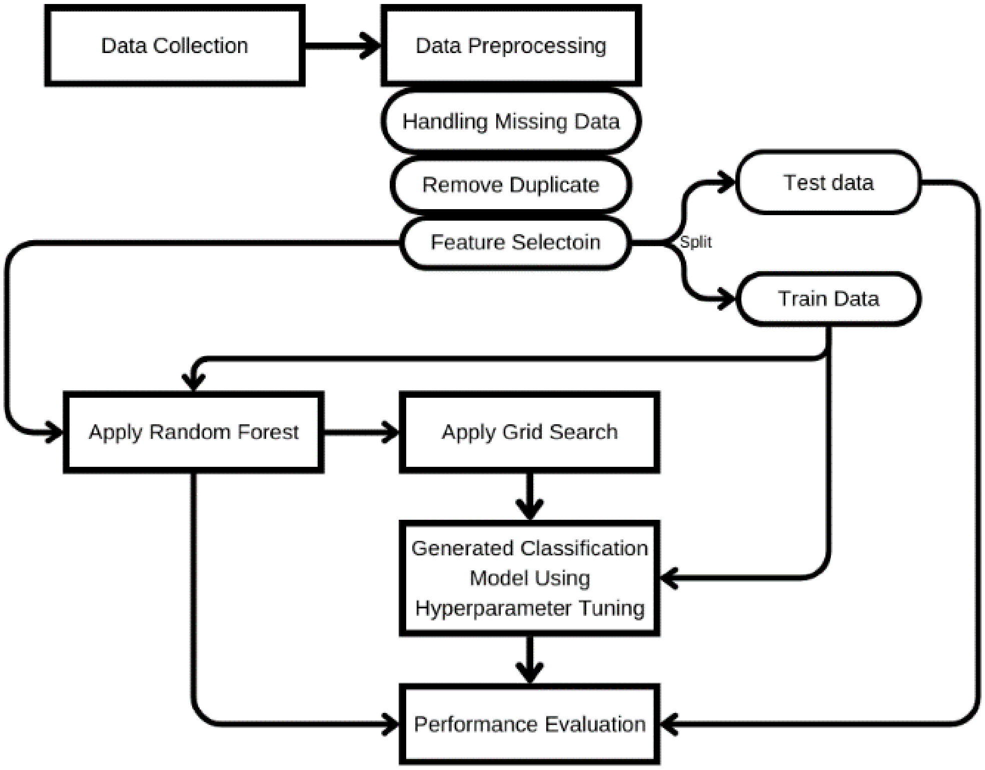


Figure 3.3: Workflow Diagram of Cancer Detection

3.4 Data Collection Procedure

Wisconsin Breast Cancer Orginal [1] The primary dataset for cancer detection is an original dataset obtained from the UCI Machine Learning Repository. It has 699 instances, 9 features, and 1 target class which is benign or malignant. Where the number of benign is 458 and the number of malignant is 241. It has null values in the dataset also.

3.5 Proposed Methodology

3.5.1 Random Forest

A well-liked supervised machine learning algorithm for classification and regression issues in machine learning is the Random Forest Algorithm. As we all know, a forest is made up of many trees, and the more trees it has, the more robust it is. Similarly, a Random Forest Algorithm's accuracy and capacity for solving problems increase with the total number of trees in the algorithm. A Random Forest classifier uses multiple decision trees on different dataset subsets and averages them to increase the dataset's predictive accuracy. Its foundation is the idea of ensemble learning, which is the process of merging several classifiers to solve a challenging issue and enhance the model's functionality [11].

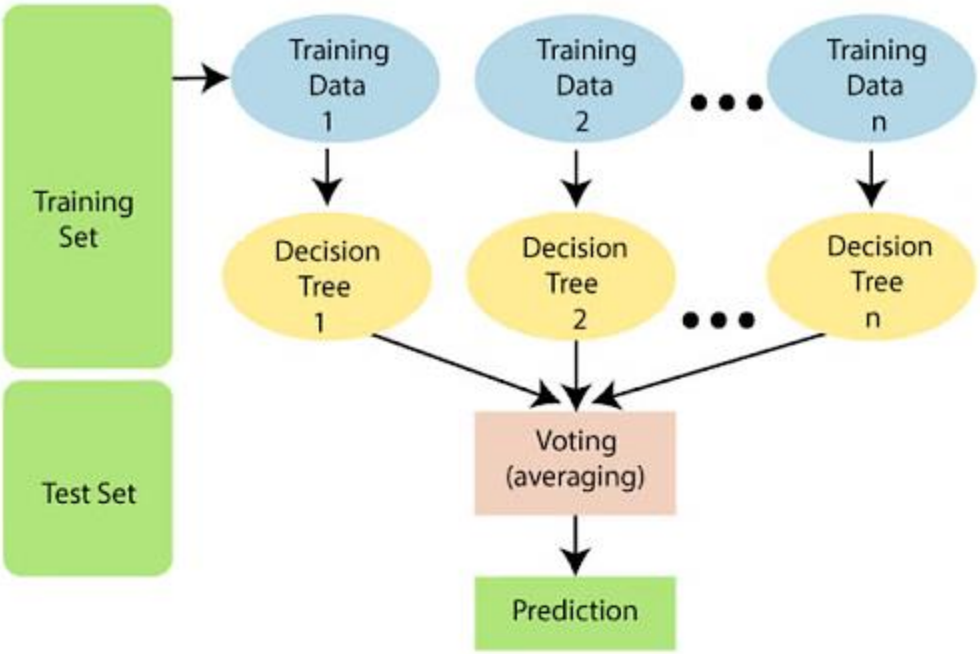


Figure 3.5.1: Working of Random Forest Algorithm [11].

The Random Forest Algorithm's operation is explained in the steps that follow:

Step 1: From a specific training set or data, choose samples that are random.

Step 2: For each training set of data, this technique will build a decision tree.

Step 3: The decision tree is going to be averaged to determine the best.

Step 4: Choose the prediction result that received the most votes to be the final outcome.

3.5.2 Hyperparameter Optimization

The performance of machine learning models is greatly impacted by hyperparameters, which are different values of parameters that are used to regulate the process of learning. The maximum depth (max_depth), criteria, and number of estimators (n_estimators) in the Random Forest algorithm. These adjustable parameters have a direct impact on a model's training efficiency. The process of determining the perfect combination of hyperparameter values to obtain maximum performance on the data in a reasonable amount of time is known as hyperparameter optimization. This procedure is essential to a machine learning algorithm's ability to make accurate predictions. Hyperparameter tuning is therefore thought to be the most challenging aspect of creating machine learning models. The hyperparameters of the majority of these algorithms used for machine learning are set to their default values. However, on various kinds of machine learning tasks, the standard values don't always work well. For this reason, must optimize them to find the ideal combination that will provide the best results [12].

In hyperparameter optimization (HPO), multi-objective optimization is an effective approach that seeks to achieve a harmonious balance between several conflicting objectives. Traditional single-objective optimization fails to capture the complex trade-offs necessary for optimal solutions because different objectives frequently conflict with one another. Introducing multi-objective optimization, which gives HPO more power by taking into account several goals at once, including model interpretability, computational efficiency, accuracy, and fairness. Multi-objective HPO traverses the hyperparameter landscape by utilizing advanced methods like evolutionary algorithms and Bayesian optimization or Grid Search. This enables it to explore a variety of options and gives decision-makers an extensive choice of options. This helps both researchers and practitioners make well-informed decisions, revealing the fine line between conflicting goals and opening the door to more advanced machine learning models that satisfy practical needs [13].

3.5.3 Grid Search

The grid search approach is the easiest to use and understand, but it is useless when there are a lot of parameters and they are not constrained by H0. Let * = (1, 2,..., m) correspond to the perturbation parameter set that maximizes p value. Setting up a lattice search is as easy as giving a set of lower bounds vector a = (a1, a2, a3, ..., am) as well as an upper bound b = (b1, b2, b3,....., bm) for each component is by definition. In a grid search, each interval of the form [ai, bi] has n evenly spaced points, and the values ai and bi are contained within them. This results in a total of nm grid points being checked. The results of the calculations are then used to determine between each pair of points which value is highest. [14]

3.6 Implementation Requirement

A thorough examination of the necessary theoretical or statistical concepts and procedures has resulted in the creation of list requirements for such a mutation classification project. It's highly likely that the following elements will be needed:

- Compatible Software or Hardware Environment
- OS (Linux distro/Windows 7,8 or above others)
- RAM (<4 GB)
- Permanent Storage

Tools for developing models

- Python Compiler
- Jupyter Notebook

CHAPTER 4

EXPRIMENT RESULTS AND DISCUSSION

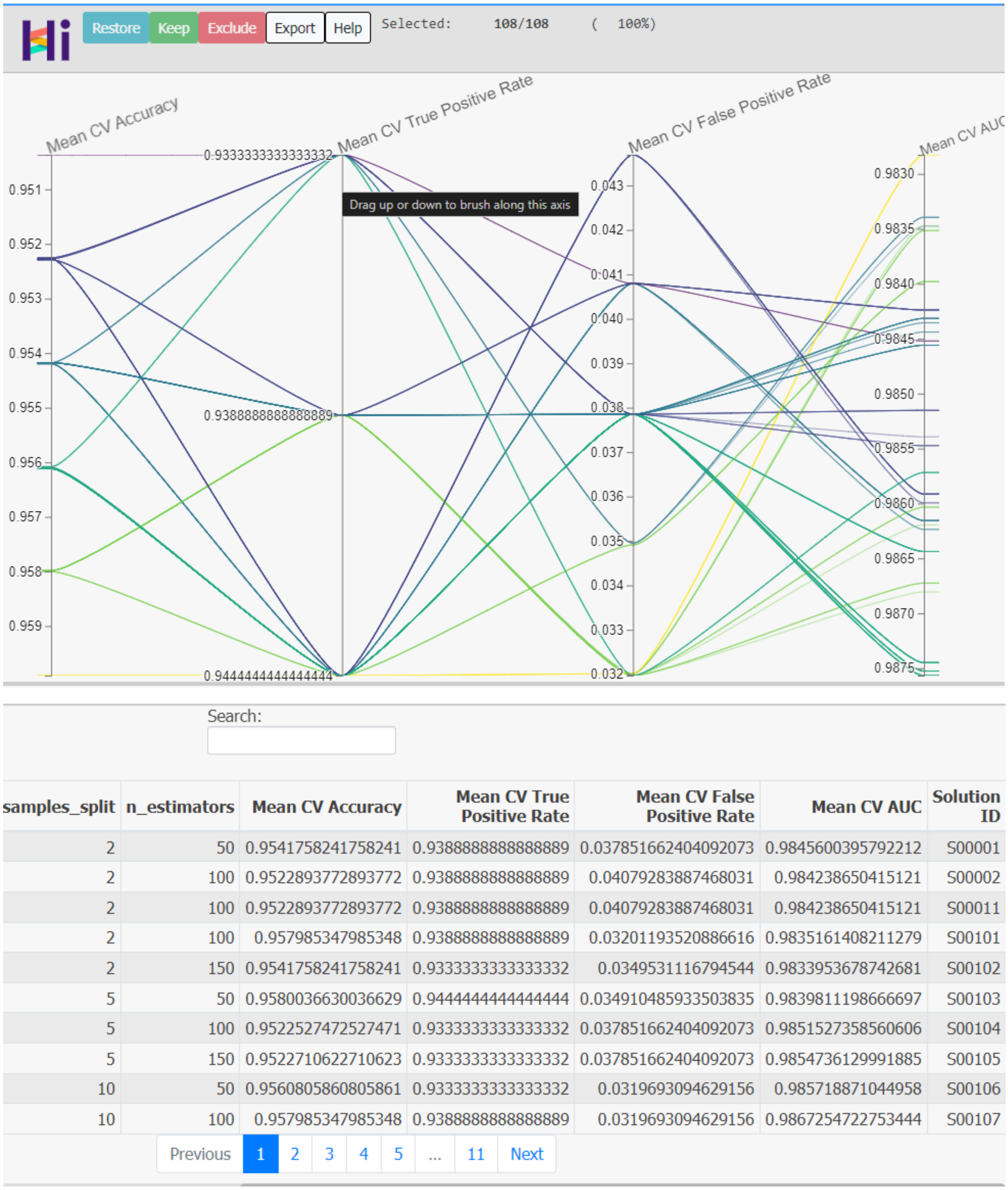
4.1 Experimental Setup

Here, output is discussed with parallel plots using the HitPlot package. The best output and hyperparameter optimized using RF and Grid Search.

4.2 Performance Analysis

First I used Default hyperparameters so the results came with Train Accuracy 1.00 and the Test Accuracy is 0.9657. That means that used Random Forest is being overfit to the training data. So here are tuned hyperparameters so as not to overfit.

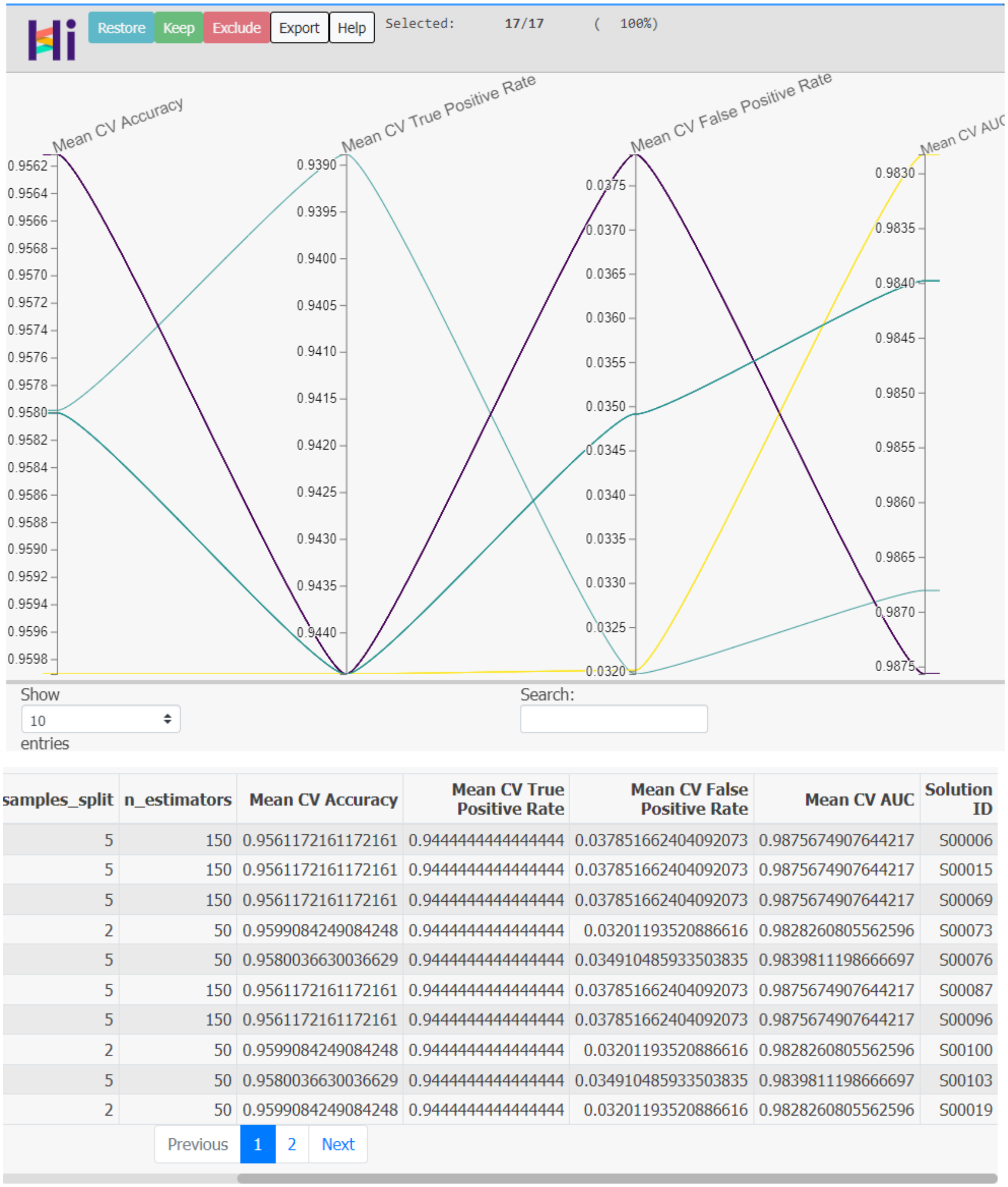
First used Single objective hyperparameter tuning that focuses on single objective which maximizes accuracy. I evaluated performance using five fold cross validation. So after performing tasks, our training accuracy decreased to 0.96 from 1 but test accuracy remain the same (0.97) which is close to test accuracy so I overcame the overfit issues. Now performing multi-objective (accuracy, true positive rate, false positive rate, and AUC) hyperparameter tuning. So I begin by calculating each of those objectives for each of our grid's 108 hyperparameter combinations. These objectives are assessed using a five-fold cross-validation method on the training set, just like in the single objective case. Then I easily plot it using parallel plots by HitPlot. Here it looks like in Figure 4.2(a)



samples_split	n_estimators	Mean CV Accuracy	Mean CV True Positive Rate	Mean CV False Positive Rate	Mean CV AUC	Solution ID
2	50	0.9541758241758241	0.9388888888888889	0.037851662404092073	0.9845600395792212	S00001
2	100	0.9522893772893772	0.9388888888888889	0.04079283887468031	0.984238650415121	S00002
2	100	0.9522893772893772	0.9388888888888889	0.04079283887468031	0.984238650415121	S00011
2	100	0.957985347985348	0.9388888888888889	0.03201193520886616	0.9835161408211279	S00101
2	150	0.9541758241758241	0.9333333333333332	0.034953116794544	0.9833953678742681	S00102
5	50	0.9580036630036629	0.9444444444444444	0.034910485933503835	0.9839811198666697	S00103
5	100	0.9522527472527471	0.9333333333333332	0.037851662404092073	0.9851527358560606	S00104
5	150	0.9522710622710623	0.9333333333333332	0.037851662404092073	0.9854736129991885	S00105
10	50	0.9560805860805861	0.9333333333333332	0.0319693094629156	0.985718871044958	S00106
10	100	0.957985347985348	0.9388888888888889	0.0319693094629156	0.9867254722753444	S00107

Figure 4.2(a): Output (Dominated)

Each hyperparameter combination is represented as a single line on this plot. The solution that achieved the best results on all objectives would be a straight line across the bottom, and this is how we see the objectives trading off performance with one another (lines crossing). We oriented the axes so that down is optimal. So here we can see many non dominated A solution must perform the same or higher on all objectives and strictly higher on at least one in order to "dominate" others. Following this line of reasoning, the nondominated solutions the opposite of dominated are the only ones that are truly important. Therefore, I used a nondominated sort. After non dominated sort the output:



Selected: 17/17 (100%)

Show 10 entries

Search:

samples_split	n_estimators	Mean CV Accuracy	Mean CV True Positive Rate	Mean CV False Positive Rate	Mean CV AUC	Solution ID
5	150	0.9561172161172161	0.9444444444444444	0.037851662404092073	0.9875674907644217	S00006
5	150	0.9561172161172161	0.9444444444444444	0.037851662404092073	0.9875674907644217	S00015
5	150	0.9561172161172161	0.9444444444444444	0.037851662404092073	0.9875674907644217	S00069
2	50	0.9599084249084248	0.9444444444444444	0.03201193520886616	0.9828260805562596	S00073
5	50	0.9580036630036629	0.9444444444444444	0.034910485933503835	0.9839811198666697	S00076
5	150	0.9561172161172161	0.9444444444444444	0.037851662404092073	0.9875674907644217	S00087
5	150	0.9561172161172161	0.9444444444444444	0.037851662404092073	0.9875674907644217	S00096
2	50	0.9599084249084248	0.9444444444444444	0.03201193520886616	0.9828260805562596	S00100
5	50	0.9580036630036629	0.9444444444444444	0.034910485933503835	0.9839811198666697	S00103
2	50	0.9599084249084248	0.9444444444444444	0.03201193520886616	0.9828260805562596	S00019

Figure 4.2(b): Output (Non-Dominated)

we see that all solutions filtered out to just 17 non dominated solutions. These are the all optimal solutions. This is called Pareto optimality. And here can see accuracy and the false positive rate are redundant objectives. This implies that in addition to increasing accuracy, I also minimized the number of patients we misdiagnose as having cancer. We observe that true positive rate and accuracy are typically in conflict. This indicates that our model is scaring people by telling them they have cancer when they don't in order to maximize accuracy. This multi-objective approach has the advantage of allowing us to visually see the extent to which these objectives trade off. Here solution S00019 can be picked because its F1 score is the highest or if we want mainly AUC then S00006 is the best optimal solution from all combinations. And at last, I tested that the solution has good cross validated accuracy and also has good accuracy on test data by ranking the test performance of each objective and comparing the test performance to thee cross validation performance.

4.3 Result Discussion

If we look at the curve can see S00019 or the yellow curve is the optimal solution counting on F1 Score(0.96) but S00006 or Purple curve performed best counting on AUC(0.98) and we can see the other trade off. Furthermore, a solution's ability to achieve good cross-validation accuracy is verified by ranking test results for each objective and contrasting test results with cross-validated results. So here is the overall result table:

Table: 4.3: Table of Results.

Models	**Performance**
Random Forest with Default Hyperparameter	Train Accuracy: 1.00 Test Accuracy: 0.9714
Random Forest with hyperparameter tuning and cross validation (Single objective focused on accuracy)	Train Accuracy: 0.96 Test Accuracy: 0.97
Random Forest with hyperparameter tuning (Multi-objective)	Found 17 different optimal solution

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTANABILITY

5.1 Impact on society

There is great potential for society in the research on optimizing the Random Forest model for cancer detection using a multi-objective hyperparameter tuning approach. Through a balanced approach to hyperparameter tuning and addressing the problem of overfitting, the developed model shows increased accuracy in cancer diagnosis. The prospective benefits of this development to individuals include the possibility of early and precise detection, which is essential for efficient treatment planning and prognosis. In addition, the model's tuning process takes into account a number of objectives with the goal of minimizing false positives as well as false negatives, guaranteeing a more sophisticated and trustworthy diagnostic tool. Regardless of regional or demographic disparities, the model's ability to generalize across a variety of patient populations represents a step toward ensuring that everyone has fair access to accurate cancer detection. By reducing needless treatments, the optimized Random Forest model can be integrated into clinical decision support systems to empower medical professionals with improved diagnostic capabilities while also optimizing resource usage. Beyond the direct effects on healthcare, the study demonstrates the potential of cutting-edge machine learning methods in the field, spurring additional advancements at the nexus of AI and healthcare. The impact on society ultimately resides in better diagnostic precision, a decrease in medical errors, more efficient use of resources, and a wider adoption of innovative technologies in medical procedures, all of which advance healthcare and benefit patients globally.

5.2 Ethical Aspect

The responsible development of technology in healthcare is heavily reliant on the ethical implications of research on improving the Random Forest model for cancer detection. The study highlights the significance of gaining informed consent for dataset utilization and calls for a commitment to patient privacy, data security, and openness. In order to avoid unintentional discrimination, efforts must be made to reduce biases and guarantee fairness in model outcomes. Since it facilitates the clear communication of predictions to patients and healthcare professionals, the interpretability of the machine learning model is essential for accountability and trust. It is essential to conduct thorough clinical validation and responsible deployment, which includes obtaining explicit consent before using the model to treat patients. Ethical standards are reinforced over time by ongoing oversight and revisions to take changing healthcare practices and demographics into account. A reliable and moral integration of cutting-edge technologies into the healthcare sector is facilitated by open communication and collaboration between researchers, medical professionals, and the general public.

5.3 Sustainability

Research on improving the Random Forest model for cancer detection has two long-term sustainabilitys. First off, the carbon footprint of data processing is reduced when ecologically friendly computing techniques, like algorithm optimization, are used. Second, ongoing evaluation, updates, and adaptation to evolving healthcare environments guarantee the model's long-term applicability and foster long-term advantages for both patients and healthcare systems. The study's dedication to ethical issues, such as minimizing bias and maintaining open lines of communication, is consistent with the values of justice and builds confidence. Furthermore, cooperative endeavors and information exchange among scientists support the overall sustainability of advances in machine learning and cancer detection.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Introduction

Using a multi-objective hyperparameter tuning approach, this thesis successfully improved a Random Forest model for cancer detection. Using the dataset, the study started with feature selection and used single and multi-objective hyperparameter tuning . The area under the receiver operating characteristic curve, accuracy, false positive rate, and true positive rate were all trade-offs examined in the multi-objective tuning phase that followed. Understanding intricate metric relationships through visualization helped identify nondominated solutions that strike a balance between several goals.

6.2 Conclusion

In the study, I proposed a ML model that detects breast cancers. For datasets Breast Cancer Wisconsin (Original) [1] dataset used. And I tuned hyperparameters for single and multiobjective. Then can see the tradeoff between different solutions. I used here Random Forest classifier and grid search for tuning hyperparameters and 5 folds cross validation the data. In addition to improving the model's accuracy, this study offered insights that are essential for attaining a comprehensive and well-rounded performance in cancer detection. The focus on nondominated solutions emphasizes how important it is to take into account a variety of objectives when evaluating machine learning models, especially when it comes to applications in the healthcare industry.

6.3 Future Work

Enhancing model performance in the future requires examining different ensemble learning techniques and fine-tuning the Random Forest algorithm for cancer detection. Expanding multi-objective hyperparameter tuning to include more metrics especially those unique to cancer can provide a more thorough analysis. Improved feature representation may be possible through integration with cutting-edge technologies like deep learning. The model's application in clinical settings is ensured by real-world validation employing a variety of datasets and giving interpretability top priority. In order to match the model with clinical requirements and ethical concerns, collaboration with healthcare professionals is still essential. By addressing these issues, machine learning will continue to progress, resulting in more potent cancer detection instruments for the medical field.

REFERENCES

[1] W. Wolberg, "Breast Cancer Wisconsin Orginal," UCI Machine Learning Repository, 1992. [Online]. Available: https://doi.org/10.24432/C5HP4Z. [Accessed 2024].

[2] A. S. Parampreet Kaur, "BSense: A parallel Bayesian hyperparameter optimized Stacked ensemble model for breast cancer survival prediction," *Journal of Computational Science,* 2022.

[3] G. W. A. K. Dursun Delen, "Predicting breast cancer survivability: a comparison of three data mining methods," *ELSEVIER,* pp. 113-127, 2004.

[4] A. J. Bamidele, "An Enhanced Hyper-Parameter Optimization of a Convolutional Neural Network Model for Leukemia Cancer Diagnosis in a Smart Healthcare System," *MDPI,* p. 2, 2022.

[5] Z. T. P. P. M. C. Mamta Jadhav, "Breast Cancer Prediction using Supervised Machine Learning Algorithms," *nternational Research Journal of Engineering and Technology (IRJET),* vol. 06, no. 10, pp. 851-853, 2019.

[6] C. R. Wolfe. [Online]. Available: https://towardsdatascience.com/training-a-random-forest-to-identify-malignant-breast-cancer-tumors-49e8a69fc964. [Accessed 7 1 2024].

[7] S. E. F. Mohammed Amine Naji, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *iNTERNATIONAL WROKSHOP ON EDGE IA-IOT FOR SMART AGRICULTURE,* pp. 487-493, 2021.

[8] A. T. Azar, "Decision tree classifiers for automated medical diagnosis," *NEURAL COMPUTING AND APPLICATIONS,* pp. 2387-2403, 2013.

[9] M. S. M. Desai, "An anatomization on Breast Cancer Detection and Diagnosis employing Multi-layer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN)," *ClinicaleHealth* , 2020.

[10] R. R. Satish Chaurasiya, "Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification," *Wireless Personal Communications* , pp. 1-10, 2023.

[11] "Random Forest Algorithm," Simplilearn, 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm. [Accessed 2024].

[12] D. David, "Hyperparameter Optimization Techniques to Improve Your Machine Learning Model's Performance," freecodecamp, 2020. [Online]. Available: https://www.freecodecamp.org/news/hyperparameter-optimization-techniques-machine-learning/. [Accessed 2023].

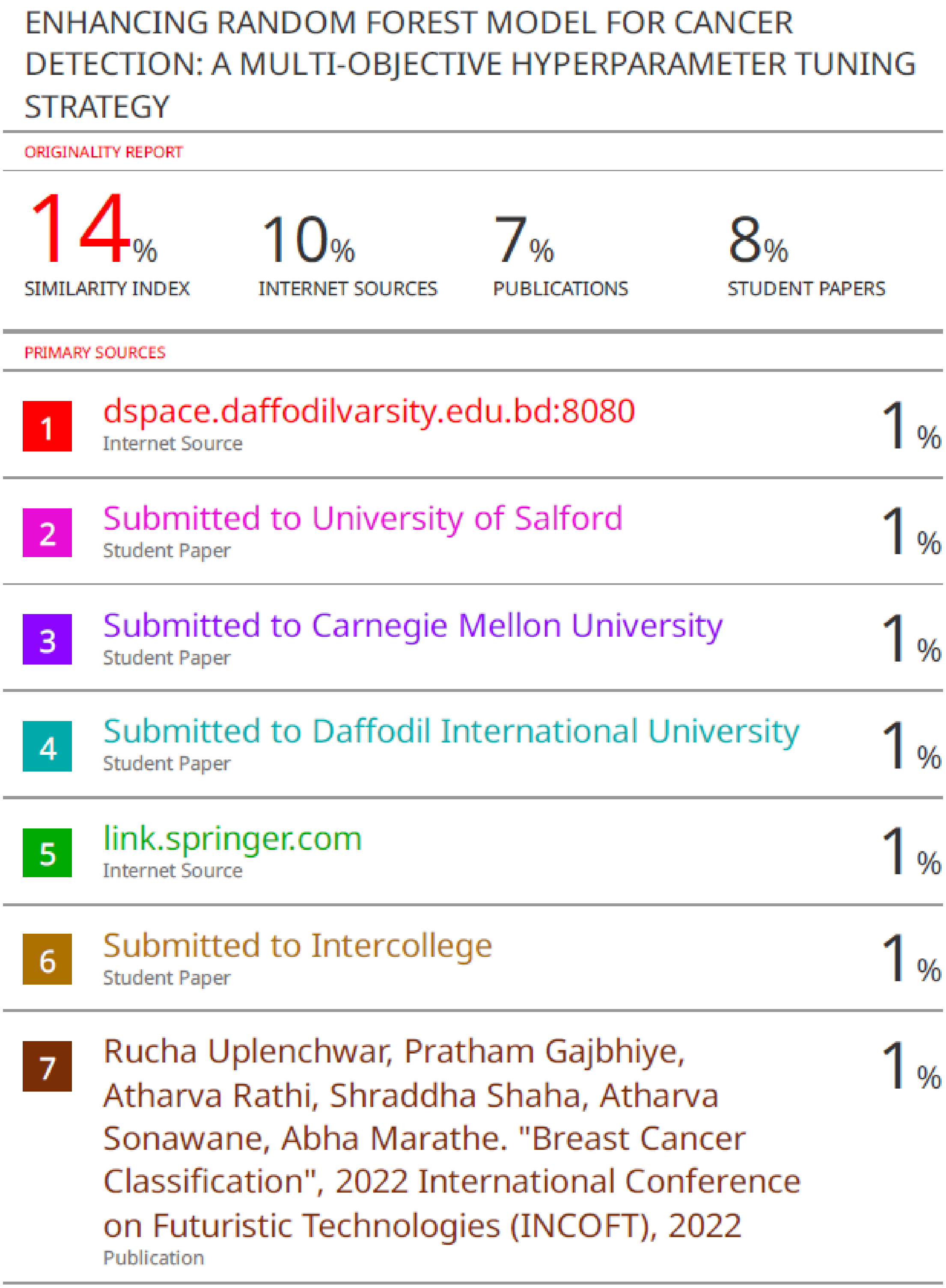
[13] "automl.org," automl.org, 2023. [Online]. Available: https://www.automl.org/dl-2-0/multi-objective/#:~:text=Multi%2Dobjective%20optimization%20in%20hyperparameter,balance%20between%20multiple%20competing%20objectives...

[14] W. KOEHRSEN, "Intro to Model Tuning: Grid and Random Search," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/code/willkoehrsen/intro-to-model-tuning-grid-and-random-search. [Accessed 2023].

APPENDICES

While completing my research I faced numerous difficulties. The first was to identify the algorithm and determine the projects methodologicals approach. However, the biggest-difficulty I faced during research was a lack of research understanding. However, I overcame it by studying and analyzing related sectors. My next task was to find a more accurate algorithm for the projects.

PLAGIARISM REPORT



ENHANCING RANDOM FOREST MODEL FOR CANCER DETECTION: A MULTI-OBJECTIVE HYPERPARAMETER TUNING STRATEGY

ORIGINALITY REPORT

14%	10%	7%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1 %
2	Submitted to University of Salford Student Paper	1 %
3	Submitted to Carnegie Mellon University Student Paper	1 %
4	Submitted to Daffodil International University Student Paper	1 %
5	link.springer.com Internet Source	1 %
6	Submitted to Intercollege Student Paper	1 %
7	Rucha Uplenchwar, Pratham Gajbhiye, Atharva Rathi, Shraddha Shaha, Atharva Sonawane, Abha Marathe. "Breast Cancer Classification", 2022 International Conference on Futuristic Technologies (INCOFT), 2022 Publication	1 %