Weighted-Based and Range-Based Breast Cancer Prediction Model using Machine Learning, Deep Learning and Fusion Models

Thesis of the Ph.D. Dissertation

Sam Khozama

Scientific Advisers:

Zoltán Nagy, Ph.D.

Zoltán Gáspári, Ph.D.

Pázmány Péter Catholic University

Faculty of Information Technology and Bionics

Roska Tamás Doctoral School of Sciences and Technology

2024

Contents

Figures

Tables

List of abbreviations

Abbreviation	Explanation
AUC	Area under curve
BCIMS	Breast Cancer Information Management System dataset
BCSC	(Breast Cancer Surveillance Consortium) dataset
BCWD	Breast Cancer Wisconsin Diagnostic
Bi-LSTM	Bidirectional LSTM
BMI	Body Mass Index
BNs	Bayesian networks
CNN	Convolutional Neural Networks
DL	Deep Learning
DNN	Deep neural network
DOI	Degree of importance
DT	Decision Trees
DTR	Decision Trees Regression
ELM	Extreme Learning Model
FDR	False discovery rate
FNR	False negative rate
HNSCC	Head and Neck Squamous Cell Carcinoma
HN-PET-CT	Head-Neck-PET-CT
K-NN	K-Nearest neighbor
LR	Logistic Regression

LSTM	Long-short term memory
MCE	Minimum Classification Error
ML	Machine Learning
MLP	Multi-layer perceptron
NB	Naïve Bayes
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
OPC	Oropharynx Cancer
ORB	Örebro dataset
PPR	Positive predictive rate
RF	Random Forests
RFR	Random Forest Regression
RNN	Recurrent Neural Network
SVM	Support vector machines
TPR	True positive rate
UQ	Uncertainty Quantification
WBCD	Wisconsin Breast Cancer Database
WHO	World Health Organization

This thesis is dedicated to my parents for their boundless support, and to the great philosophers, including Pythagoras, who enriched our understanding of philia—love and friendship—and whose ideas continue to inspire and guide my intellectual journey.

Acknowledgments

I sincerely appreciate everything my supervisors, Péter Szolgay, Zoltán Nagy, Zoltán Gáspári, and Ali Mayya, have done for me throughout my Ph.D. journey. Their guidance and insightful comments were extremely helpful in determining the focus and outcome of this dissertation.

I would like to extend my deepest gratitude to my family and friends for their constant encouragement, inspiration, and comprehension throughout this challenging endeavor.

Sam Khozama

Abstract

Breast cancer prediction is a challenging area of medical engineering. In this study, I presented a novel breast cancer range-based and weighted-based prediction model based on machine learning and deep learning algorithms. However, the study branched into five branches. In the first one, a novel weighting algorithm was proposed and applied on the well-known BCSC dataset. In the second branch, a novel range-based breast cancer prediction model was proposed. While in the third branch, a fusion model of the best ensemble ML and DL models was designed and evaluated. The probabilistic model was applied again in the fourth fork but on the whole dataset computing the new distribution of the target column without any balancing operations. Three regression models are proposed, in the fifth and last section, using again the whole dataset.

Within the initial branch of this research, I examined the impact of risk factor weighting and selection, along with testing three versions of a balanced dataset. The experiments conducted demonstrated that the weighting technique considerably improved accuracy and decreased errors. The overall test accuracy was 95.8%.

In the second branch, a novel range-based breast cancer prediction model was introduced. The BCSC dataset was analyzed using a probabilistic model to determine the final prediction value for each case in the dataset.

This new score was used to update the BCSC dataset, and the resulting modified dataset was then utilized to train an ensemble learning model using the Bayesian hyperparameters optimization method.

The training process was conducted in two scenarios, one using the entire dataset and the other using a subset consisting of 67,633 samples. In both scenarios, the MCE, TPR, PPR, and FDR were computed in three cases, the first being the 0-variance scenario in which no error margin was allowed, while in the second and third cases, ±1 class-variance tolerance was applied (since very closely related subclasses yield similar results).

The results demonstrated improvements in TPR, PPR, and accuracy for the (±1 and ±2) variance scenarios for both the sub-dataset and entire dataset. Additionally, the new modified version of the BCSC dataset contained more detailed information about breast cancer prediction than the old version, which only indicated the presence of cancer without any

percentage. The new version of the BCSC dataset was available as supplementary material for future research.

In the third branch, I proposed a range-based breast cancer prediction system that combines two different models: an LSTM deep learning model and an ensemble of boosted decision trees.

Initially, I used a balanced range based BCSC dataset and preprocessed it by categorizing the classes into seven different categories.

The new dataset was then split into training and testing sets, with a 25% ratio for the test set. The DL model consisted of a sequence input layer, one LSTM layer, one fully connected layer, and one SoftMax-classification layer. The ML model, on the other hand, comprised 30 decision tree classifiers trained using hyperparameters optimization and boosted learning approach. The final prediction was obtained by fusing the ML and DL scores.

I performed experiments using five different training scenarios, involving different LSTM cells and training epochs.

The fourth and fifth branch are performed on the original dataset but after applying the probabilistic model to get the target column in its range-based score.

The fourth branch is a classification task by which some of the result categories of the target column (prediction score) is merged to minimize the difference between number of samples of these categories. The fifth branch is performed as a regression task in which the target column is treated as a continuous column. In both fourth and fifth branches, many ML and DL classification and regression models are trained and evaluated. The ensemble of ML and DL models are also used to improve the performance. All models are evaluated using many classification and regression metrics besides many statistical and medical analysis tools like Violin plots, variance plots and distribution plot.

Results of fourth and fifth branches prove the high accuracy of the predicted range-based score in both classification and regression models.

Keywords:

Breast Cancer, Machine Learning, Deep Learning, Risk Factors, Cancer Prediction, BCSC Dataset, Risk Prediction, Regression Models.

Chapter 1
Introduction

1.1. Introduction

Data analysis has become one of the fastest-growing fields in computer science, owing to the exponential growth in dataset sizes. Cancer prediction is among the fields benefiting from data analysis and Machine Learning (ML) algorithms for estimating cancer risk [1] [2] [3].

The use of ML techniques can enhance the performance of cancer prediction, leading to a significant improvement (15%-20%) in estimation accuracy over the past few years [4]. Breast cancer prediction can identify potentially high-risk women and guide them towards improving their lifestyle, thereby avoiding future therapy and costs [5].

Known risk factors participate in causing half of cancer cases [6] [7]. Several risk factors contribute to breast cancer, including early menarche, late menopause, obesity, age at first birth, and hormone therapy, which impact the exposure duration of breast tissue to hormones that increase the risk of developing cancer [8] [9].

The primary challenge in cancer diagnosis and prediction is the vast amount of data that cannot be managed using traditional manual methods (such as physician's observations). A more efficient and rapid approach is required to address this issue [10] [11] [12] [13].

Thankfully, the rapid advancements in the field of computer science have enabled the extraction of valuable insights from large datasets, equipping healthcare organizations with effective tools for diagnosing and predicting cancer [6] [14] [15] [16].

1.2. Research Background

Several breast cancer prediction systems have been developed in the past, utilizing various machine learning models such as Support Vector Machines (SVM), K-Nearest Neighbor (K-NN), Random Forests (RF), Decision Trees (DT), Neural Networks, Naïve Bayes (NB), and Logistic Regression (LR) [17] [18] [19] [20] [21].

Some researchers have integrated deep learning techniques with image models to extract mammography breast image features and text-based risk factor information, resulting in improved prediction model accuracy [22].

Parameter optimization and ensemble learning methods have also been employed in some studies to significantly enhance the system's performance [21] [23] [24].

1.3. Research Importance

Breast cancer is one of the most common types of cancer among women worldwide. The early prediction and detection of this disease can help in the therapy and safe lives. Machine learning algorithms have shown good results in predicting and detecting breast cancer. Such tools can assist the medical professionals and physicians in making better decisions.

The importance of our research can be concluded in the following:

1. Early detection: Machine learning algorithms can analyze and process big amounts of patient data and identify patterns that may not be apparent to human experts. By doing so, they can help physicians to identify breast cancer at an early stage when treatment is most effective.
2. Personalized treatment: Machine learning algorithms can determine the most effective treatment plan for each patient based on their unique characteristics, such as age, medical history, and genetic information.
3. Improved accuracy: Machine learning algorithms can analyze complex data sets and define changes that may not be noticeable to humans. This can lead to more accurate and reliable diagnoses.
4. Cost-effective: By using machine learning algorithms to analyze cancer data, medical professionals can reduce the need for costly and invasive diagnostic tests, such as biopsies (especially in case of breast cancer).

1.4. Open-Ended Questions

Previous studies on cancer prediction have utilized popular machine learning models, while only a few have explored the concept of ensemble learning or combining multiple models. Several studies have employed the BCSC (Breast Cancer Surveillance Consortium) dataset, either partially or in its entirety, but none of them have analyzed the dataset's probabilistic distribution. Previous studies have approached the cancer prediction problem by providing specific binary results (yes or no). In our study, we aim to compute a range-based cancer score using a probabilistic model.

Here are the main research questions:

How can we obtain a range-based breast cancer score?

Which dataset is suitable for predicting breast cancer?

What are the essential risk factors that are best-predicting breast cancer?

What is the best approach to selecting the best combination of risk factors?

How can we use machine learning and deep learning methods to predict breast cancer based on weighted selected risk factors?

What is the best way to deal with unbalanced breast cancer datasets?

How classification and regression differ in the problem of range-based breast cancer prediction?

1.5. Research Aims

The main aim of this study is to build a range-based breast cancer prediction system based on machine learning algorithms. Three parts are included in our work, including three specific objectives. In the first part, our focus is on selecting the best combination of risk factors by using a weighting methodology assigning a degree of importance to each risk factor. This part aims to guide the importance of each risk factor in the final prediction score (the more essential the risk factor, the more degree of importance). In the second part, a range-based breast cancer prediction is our objective. This part aims to make the cancer prediction technique predict risk with a percentage and not only (0/1) values. The final part aims to use the well-known LSTM deep learning architecture in the prediction of breast cancer in order to enhance performance.

1.6. Scope of Research

The breast cancer prediction system needs two specific tools; a good statistical dataset and a suitable prediction artificial intelligence approaches. To achieve our goals, three branches are involved in the current study:

The first state-of-art of breast cancer prediction is based on assigning a weight value to each risk factor based on their importance. This mechanism allows us to select the most essential risk factors and provide them to the machine learning models.

The second branch introduces a novel range-based breast cancer prediction system based on the weighted selected risk factors of the first branch. The model also uses the weighting methodology to achieve the best fusion of the BCSC's risk factors.

In the third part of this study, we developed a fusion model of two machine learning and deep learning models. To obtain the final prediction, Long-Short Term Memory (LSTM) and ensemble learning with hyper parameters optimization are used, and score-level fusion is used.

1.7 Workflow

The study starts with studying many pieces of research in the field of breast cancer prediction. The main limitations of these studies are summarized, and the novel state-of-art of the current research is clarified and organized. BCSC dataset is used as the main risk factor dataset. The dataset is not balanced so it needs some preprocessing steps before proceeding with the prediction part. In the balancing step, we suggest using three different balancing approaches, including the over-sampling, the down-sampling, and the mixed approach. In the next step, a novel methodology is proposed to define the degree of importance of each risk factor in order to select the most appropriate risk factors for the next prediction step. The method is based on many medical questionnaires and a statistical study of the most recent medical studies and related medical datasets. After defining the degree of importance of each risk factor, many training scenarios can be used to define the best combination of risk factors.

The next part of the study introduced a novel range-based breast cancer prediction tool depending on giving a range-based score and not only (0/1) score. This part includes using the balanced version of the BCSC dataset and the weighting and selection mechanism of the first part. The new method depends on different statistics (previous medical knowledge, the likelihood of each risk factor given all prediction classes, cancer probabilities and non-cancer probabilities). The final prediction score is computed using the post-probability of the weighted combination of risk factors and the acquired statistical probabilistic model. In the next step, an ensemble learning model is suggested to achieve optimal performance. For the third part of this study, a fusion of machine learning and deep learning methods is proposed. Additionally, the outputs of the first two sections of this study could serve as inputs or auxiliary methods for the third part.

The fourth and the fifth parts of this study included studying the original dataset with applying of the probabilistic model derived from the second part of the study. For the fourth part, the target column of the dataset is transformed into a range-based score and the adjacent categories are merged, while in the fifth part a regression task was performed so the target column is only

transformed into range-based scores without any merging. The dataset was then split into training and validation, and the SMOTE balancing algorithm was only applied to the training set. Many ML and DL classification and regression models were applied within those two parts. The ensemble learning of the best ML and DL models was also utilized.

Classification and regression-based Evaluation metrics besides the statistical and medical analysis were all used in the evaluation process of the fourth and fifth branches.

Chapter 2

Literature review

2.1. Introduction

Breast cancer is classified as one of the most common cancer types [25]. According to the World Health Organization (WHO), cancer is the second leading cause of death [26] [27]. Breast and oral cavity cancers are considered the causes of 25% of deaths around the world [25].

Based on cancer statistics from 2020, breast cancer constitutes 11.7% of all cancer records around the world [28].

From the death side, breast cancer was classified as the second deadliest cancer after lung cancer by a percentage of 6.9% [28].

All these previous facts lead to the importance of the prediction of breast cancer before actual diagnosis. Early prediction can reduce the cancer rate and help physicians predict cancer at its early stages. Fortunately, computer science algorithms have been incrementally developed and enhanced and can be used for the purpose of cancer prediction. Physicians themselves cannot process and analyze all the cancer data since it is huge and very related. Consequently, they need the efficiency of computer science algorithms that can handle large amounts of data in a short time.

Many previous systems had been introduced in the field of breast cancer prediction. Some of them used the logistic regression approaches [29] [30] [31] [32], while others used the neural networks [29] [31] [33]. Other data mining algorithms like decision trees [29] [34], Naïve Bayes methods [34], Support Vector Machines [30] [31] [35], Random Forests (RF) [30] [31] [32], optimization algorithms [35], etc.

Many datasets were used for breast cancer estimation. The Breast Cancer Surveillance Consortium dataset [36] is one of the most common datasets. It consists of 2,392,998 screening mammograms, 280,660 records and 13 risk factors. This dataset had been used in many pieces of research [34] [37]. Another international dataset is the Breast Cancer Information Management System (BCIMS) dataset consisting of 16,000 cases [38] and was used by studies

[39] [40]. Some other researchers collected their datasets from specialized medical centers or hospitals [41] [42].

2.2. Related Work

Breast cancer prediction have received a good attention in the scientific researches. Many research centers and international institutions introduced papers in the field of cancer prevention and detection. In this section, I will introduce the most recent studies in this field with a comparative and analytical discussion.

I will split the studies into two main parts; the first one deals with the general studies on breast cancer prediction, while the second one concentrates on the probabilistic-based approaches (since our second part of the study deals with the range-based cancer prediction).

2.2.1. General breast cancer prediction methods

Shieh et al. [43] proposed a breast cancer prediction model using the information of the clinical and polygenic risks. The Bayes estimation and conditional logistic regression models are used together to study the common effect of ordinary and polygenic risk factors on the future risk of breast cancer. The researchers used 486 cases of the BCSC dataset and found that prediction accuracy increased from Area under curve (AUC)=0.62 to AUC=0.65 after adding the polygenic risk to the model. They concluded that 18% of cases were classified as high-risk cases in the common model while it was only 7% for the ordinary risk factors model.

In 2020, Rajendran [34] and others used the supervised machine learning algorithms on class imbalanced data for the prediction of breast cancer on the BCSC dataset. In order to apply balancing, they used three approaches: Synthetic Minority Oversampling, under-sampling (Spread Subsample) and fusion of both techniques. They also used Bayes classifier, Bayes networks, Random Forests (RF) and random trees as classifiers. The best accuracy they obtained was 99.1% under FP (False Positive) equals 21%. The problem with research was that they used only 10,252 instances after applying the balancing techniques. The results also indicated a low sensitivity of 78.1% (low positive rate) while the accuracy was 99.1% (conflicting results).

Li and Sundararajan [44] applied several machine learning approaches for the prediction of breast cancer. They used only 10000 cases and 8 risk factors of the BCSC dataset (menopause,

age, breast density, Body Mass Index (BMI), race, first birth age, number of first-degree relatives having breast cancer and hormone therapy). SVM and Bayes classifiers were used for the final risk estimation. They got 96.6% and 91.26% as accuracy for SVM and Bayes classification respectively.

Hou et al. [33] introduced a model for the prediction of breast cancer of Chinese women using machine learning algorithms. They used 7127 cases of the BCIMS dataset. They chose specific risk factors based on the fact that they must be known and collected by the same measurement techniques. Consequently, 10 risk factors had been chosen and different prediction models were used like RF, deep neural networks DNN and XGBoost. They got an accuracy of 72.8 for both DNN and RF, while the XGBoost accuracy was 74.2%.

Kakileti et al. [32] evaluated the performance of many machine learning classifiers for the prediction of breast cancer risk under incomplete datasets. They evaluated the RF, Logistic Regression (LR) and custom Neural Network (NN). The Area Under Curve (AUC) was used for the performance evaluation. The entire BCSC dataset was divided into 75% for training and 25% for test. AUC achieved 0.645, 0.634 and 0.649 for LR, RF and NN respectively. The custom NN achieved a better performance in the case that less than 50% of the dataset was missing.

Ming et al. [45] collected a breast cancer prediction dataset from Geneva University Hospitals. Their dataset included 112587 individuals and 14 variables. They applied different ML algorithms like Markov mixed model, adaptive boosting, and RF. They obtained accuracy between 84.3% and 88.9%. However, the dataset variables related not only to breast cancer but also to other tissues. The study needs more risk factors including hormone therapy, mammographic information, other body indexes, etc.

Lang et al. [46] predicted oropharyngeal cancer using 3D Convolutional Neural Networks (3D CNN) on a dataset consisting of 675 breast cancer cases. They split the dataset into training (412 cases of the Oropharynx Cancer (OPC) dataset and 263 cases of the Head and Neck Squamous Cell Carcinoma (HNSCC) dataset) and validation (90 cases of the Head-Neck-PET-CT (HN PET-CT) dataset). For the test, they used 80 cases from the HN1 dataset. The experiments showed that the Area Under Curve (AUC) was 0.81.

In 2022, Ashokkumar et al. [47] predicted the lymph nodes of the breast using the Kohonen self-organizing ANN. They used a dataset of 10,150 images of 850 patients. Their approach achieved 94% accuracy.

Recently, Saleh et al. [48] introduced a deep learning-based breast cancer prediction model. They used the Recurrent Neural Network (RNN) with five hidden layers and one output layer. Three feature selection models were proposed. The Breast Cancer Wisconsin Diagnostic (BCWD) dataset was used in their study. It had 30 factors (features) and one class (cancer prediction 0 or 1). The results indicated an accuracy of 95.18%.

Uddin et al. [49] used the well-known Wisconsin Breast Cancer Dataset (WBCD) dataset with different machine learning methods, including SVM, RF, K-NN, DT, NB, LR, NB, AdaBoost, Multi-layer perceptron MLP, nearest cluster classifier (NCC), and voting classifier (VC). All models were evaluated using accuracy, precision, recall and F1-score (a weighted average of precision and recall indicating the accuracy of a model). Results showed that the voting classifier achieved the best accuracy 98.77%. The size of used dataset was small (569 records).

Recently, in 2023, Botlagunta et al. [50] introduced a diagnosis and classification model of breast cancer using machine learning algorithms. Many ML algorithms were used and evaluated using accuracy, ROC, and AUC. Results indicated that the DT classifier achieved the best performance of 83% accuracy and 0.87 AUC. The used dataset was of a small size (5176 records) and the prediction was either yes or no.

2.2.2. Probabilistic-based breast cancer prediction methods

Kumar et al. [51] introduced the conditional probability aspect of Bayes theorem to predict liver cancer. Their study utilized a dataset of 20 patients from the BUPA research lab, which contained seven attributes such as Mean corpuscular volume, Alkaline phosphate, alkaline aminotransferase, aspartate aminotransferase, gamma trans peptidase, the number of half-pints equivalent to alcohol, and a selector for dataset division into training and validation. The researchers computed various probabilities using this dataset, including the probability of an individual having liver cancer and the conditional probability of a positive/negative test result given the presence or absence of the disease. To analyze the dataset and apply the NB classifier, they used the Weka tool. However, the accuracy of their results was only 50%, indicating that their methodology lacked pre-processing steps and the dataset was too small.

In their study on predicting survivability after breast cancer surgery, Al-Jawad et al. [18] employed Bayesian Network and SVM methods. Their research utilized Haberman's survival dataset, which consisted of 306 cases, including 225 confirmed cancer cases that survived for five years after the surgery. The authors used the Weka tool to apply SVM and BN classifiers and computed five statistical features (mean, median, standard deviation, maximum, and minimum values) for the three attributes of the dataset. They also calculated the correlation coefficients between pairs of features (Age and survival status: 0.067, Year and survival status: -0.00477, Positive nodes and survival status: 0.28677). However, their methodology suffered from low performance due to the use of fixed values for the optimizable parameters of SVM and BN models. Their results showed that SVM outperformed the Bayesian Network by 6.88%, achieving 73.78% and 74.77% for Recall and Precision metrics, while the BN achieved 78.22% and 64.47% for Recall and Precision, respectively. The study's main limitations were the small size of the dataset and the fixed learning parameters.

Witteveen et al. [19] conducted a study in 2018 to compare logistic regression with various Bayesian Networks (BNs). They utilized a subset of data from the Netherlands Cancer Registry, comprising 37,320 samples of women with early-stage breast cancer between 2003 and 2006. To improve the performance of the BNs architectures, the authors employed Bayesian network classifiers, correlation coefficients, constraint-based learning methodologies, and score-based learning models. AUC evaluation metrics were used to assess the different models, and an external validation set from the NCR from 2007 and 2008 (N = 12,308) was obtained to apply these validations. Although logistic regression performed better in most experiments of the sub-dataset analysis, BNs outperformed regression for SP prediction for the high and low-risk subsets. The researchers concluded that the coefficient estimators' value had no correlation with the changes in the other variables' values in the case of BNs.

Yang et al. [52] conducted a study in which they fused three different classifiers (Bayesian and Markov models, and artificial neural network) to achieve optimal efficiency. They utilized Bayesian and Markov models to establish a connection between the previous and current incidence of cancer, and the outputs of these two classifiers were fed back into the Neural Network classifier. To prepare the cancer dataset, a pre-processing step was applied, including normalization and missed data manipulation. They used twenty attributes from a dataset of 36,000 cases, including 10,500 patients with lung cancer, 13,500 with liver cancer, and 12,000

with stomach cancer. The authors partitioned the dataset into 75% training and 25% test. The experimental results demonstrated that the overall training accuracy was 73.55%, 76.07%, and 75.63% for the ANN, Markov model, and the proposed fusion methodology, respectively. For the test set, the corresponding accuracies were 68.78%, 70.63%, and 72.47%. However, the main limitation of their approach was the small F1-score, indicating that their proposed method suffered from false positive and false negative results, possibly due to the data being collected from various sources. The authors also compared their results with other classifiers, such as RF, SVM, and Extreme Learning model (ELM). While their approach surpassed ELM, the performance of SVM and RF was superior.

In a study using the BCSC dataset, breast cancer prediction was performed on 154,899 records using multiple machine learning algorithms, including LR, SVM, NB, and Bayesian Network [13]. The results indicated that the NB classifier achieved the highest accuracy in predicting the likelihood of breast cancer, while SVM and BNs had lower performance.

Another research used Next-Generation Sequencing (NGS) methodology combined with machine learning algorithms for breast cancer prediction [53]. The National Center for Biotechnology Information (NCBI) dataset was employed to extract NGS data samples from four different categories, comprising 1580 samples. The sequence features were extracted, and various machine learning classifiers, such as K-NN, SVM, NB, AdaBoost, DT, RF, and gradient boosting, were utilized. The evaluation showed that the decision tree classifier achieved the highest accuracy of 94.30%.

Savic et al. [54] conducted a study on machine learning models for predicting the quality of life for breast cancer patients. They utilized two datasets, including the BcBase early breast cancer prediction dataset and the Örebro dataset (ORB) prostate cancer dataset. The authors evaluated several machine learning algorithms, such as RF, SVM, Naïve Bayes, K-NN, and decision trees, and examined two types of models, namely centrally trained and federated models. The results demonstrated that both models accurately predicted short-term predictors, while centrally trained models outperformed federated models for long-term predictors. However, the precision and recall values were low in both models.

Guo et al. [23] introduced an MLP-based cancer prediction model. The authors utilized ensemble learning to enhance the multi-layer perceptron (MLP) classifier's performance by optimizing specific parameters, such as the number of input features, hidden

layers, neurons in each layer, and weight values. The experiments were conducted on the Wisconsin Breast Cancer Database (WBCD), and the accuracy obtained using the MLP classifier and parameter optimization was 98.79%.

Combining information from multiple models can improve prediction accuracy, which is beneficial for both healthcare providers and patients. BRCAPRO is a widely used model that predicts breast cancer risk based on family history, but it has a significant limitation of not considering non-genetic risk factors. To address this issue, Guan et al. [55] expanded BRCAPRO by integrating it with another popular model, BCRAT (Gail), which utilizes a mostly complementary set of risk factors, many of which are non-genetic. They explored two approaches for combining BRCAPRO and BCRAT: (1) modifying the penetrance functions in BRCAPRO using relative hazard estimates from BCRAT, and (2) training an ensemble model that takes predictions from both models as input. Using simulated data and data from Newton-Wellesley Hospital and the Cancer Genetics Network, they demonstrated that the combination models outperformed both BRCAPRO and BCRAT. In the Cancer Genetics Network cohort, they showed that the proposed BRCAPRO + BCRAT penetrance modification model performed comparably to IBIS, an existing model that combines detailed family history with non-genetic risk factors.

In a study by Hamedani et al. [56], they evaluated three Uncertainty quantification (UQ) models for classifying breast tumor tissue types: Mont Carlo-dropout (MCD), Bayesian Ensemble, and MCD Ensemble. To improve classification accuracy and solve the problem of limited data in the Wisconsin Diagnostic Breast Cancer (WDBC) dataset used in this research, they utilized transfer learning techniques and a pre-trained Convolutional Neural Network (DenseNet121). They compared the three proposed models based on their ability to estimate the reliability of classification using novel performance metrics designed to assess the estimated uncertainty. Quantitative and qualitative analyses demonstrated that the models exhibited high uncertainty in misclassifications, which is crucial in determining the risk of medical diagnosis errors. By utilizing these new evaluation criteria, they aim to determine when it is safe to rely on the deep neural network's output. Experiments proved that the Bayesian Ensemble model provided the most reliable results.

Leventi et al. [57] implemented a probabilistic neural network for the aim of cancer prediction. They applied many steps to reach the final predication. First, they applied the data

preprocessing, preparation and embedding. Then, the neural network was trained and the decision boundary was adapted. After that, the neural network posterior predictive check (PPC) was performed. Finally, the decision boundary is back transformed into actual feature space and reasoned. In their experiments, they used the breast Cancer Wisconsin (Diagnostic) dataset (consisting of 10 columns and 569 records). They achieved an accuracy of 95.32% on the test set. Their dataset was small, and the accuracy was low according to the nature of the dataset.

Hussain et al. [58] introduced a breast cancer prediction system based on mammography images and deep convolution networks. They applied many steps to make the final prediction. First, they extracted image features (handcraft) like texture, scale-invariant features (SIFT), morphological features, Fourier descriptors, and entropy-based features. Then, the extracted features were fed into the machine learning classifiers for the classification step. Many ML and DL models, including SVM, Naïve Bayes, DT, GoogleNet, and AlexNet. were experimented. GoogleNet model achieved the best accuracy with 99.26% and AUC of 0.998, while AlexNet achieved accuracy of 99.26% and AUC of 0.9996. Their used dataset was small.

Recently in 2023, Kayikci et al. [59] designed a breast cancer prediction system based on gated attentive multimodal deep learning by which data from different resources were combined (clinical, copy number alternation and gene expression resources). In the first step of their method, stacked features were extracted using sigmoid gated attention probabilistic model, while in the second step, the classification was performed using a combination of flatten, dense and dropout layers. They used the METABRIC dataset consisting of 1980 records. They got 0.95, 91.2%, 84.1%, and 79.8% for AUC, accuracy, precision, and recall, respectively. Their dataset was small, and the model predicted the presence or absence of breast cancer only.

Table 2.1 includes a detailed comparison between previous studies.

Table 2.1. A detailed comparison between previous studies state-of-art

Researcher + Reference	Methodology	Dataset	Main Results	Train/Test set	Main Limitations
Shieh et al. [43]	Bayes estimation, logistic regression	BCSC dataset	AUC increased from 0.62 to 0.65 after adding polygenic risk. 18% classified as	Test	Low accuracy binary prediction

			high risk in common model.		
Rajendran [34]	Supervised ML algorithms	BCSC dataset	Best accuracy: 99.1%, low sensitivity (78.1%)	Test	Small dataset (10,252 instances), conflicting results
Li and Sundararajan [44]	SVM, Bayes classifiers	BCSC dataset (10,000 cases, 8 risk factors)	SVM accuracy: 96.6%, Bayes accuracy: 91.26%	Test	They take a small part of BCSC dataset not all the dataset, binary prediction
Hou et al. [33]	RF, DNN, XGBoost	BCIMS dataset (7,127 cases)	DNN and RF accuracy: 72.8%, XGBoost accuracy: 74.2%	Test	Small dataset binary prediction
Kakileti et al. [32]	RF, LR, custom NN	BCSC dataset (75% training, 25% test)	AUC: LR: 0.645, RF: 0.634, NN: 0.649	Test	Custom NN performs better with less missing data binary prediction
Ming et al. [45]	Markov mixed model, boosting, RF	Geneva University Hospitals dataset (112,587 individuals)	Accuracy: 84.3% - 88.9%, need more risk factors	Test	Variables related to other tissues, need more risk factors
Lang et al. [46]	3D CNN	Breast cancer cases (675)	AUC: 0.81	Test	Small dataset binary prediction
Ashokkumar et al. [47]	Kohonen self-organizing ANN	Dataset of 10,150 images (850 patients)	Accuracy: 94%	Test	Limited dataset size binary prediction

Saleh et al. [48]	RNN	BCWD dataset (30 factors, 1 class)	Accuracy: 95.18%	Test	Small dataset binary prediction
Uddin et al. [49]	Various ML methods	Wisconsin Breast Cancer Dataset (569 records)	Voting classifier: 98.77% accuracy	Test	Small dataset (569 records)
Botlagunta et al. [50]	Various ML algorithms	Dataset with 5,176 records	DT classifier: 83% accuracy, 0.87 AUC	Test	Small dataset (5,176 records), binary prediction
kumar et al. [51]	Naïve Bayes classifier	BUPA research lab dataset (20 patients)	Accuracy: 50%	Test	Small dataset (20 patients), lack of pre-processing
Al-Jawad et al. [18]	Bayesian Network, SVM	Haberman's survival dataset (306 cases)	SVM recall: 73.78%, Precision: 74.77%, BN recall: 78.22%, Precision: 64.47%	Test	Small dataset (306 cases), fixed learning parameters binary prediction
Annemieke et al. [19]	Logistic regression, BNs	Netherlands Cancer Registry subset (37,320 samples)	BNs outperformed regression in some cases	-	Binary prediction
Yang et al. [52]	Bayesian and Markov models, ANN	Cancer dataset (36,000 cases)	Training accuracy: ANN: 73.55%, Markov model: 76.07%, Fusion: 75.63%	Train	False positive and false negative results, data collected from various sources binary prediction
Li et al. [13]	Logistic Regression,	BCSC dataset (154,899 records)	Naïve Bayes achieved	-	Binary prediction

	SVM, Naïve Bayes, BN		highest accuracy		
Kurian et al. [53]	Multiple ML classifiers	NCBI dataset (1,580 samples)	Decision tree achieved highest accuracy (94.30%)	Test	Binary prediction
Savic et al. [54]	RF, SVM, Naïve Bayes, K-NN, decision trees	BcBase and ORB datasets	Centrally-trained models outperformed federated models	Test	Low precision and recall values Binary prediction
Guo et al. [23]	MLP classifier	WBCD dataset	Accuracy: 98.79%	Test	Binary prediction
Guan et al. [55]	Combination of BRCAPRO and BCRAT	Simulated data, Newton-Wellesley Hospital, Cancer Genetics Network	Combination models outperformed individual models	-	Small dataset, limited non-genetic risk factors
Hamedani et al. [56]	Monte Carlo-dropout, Bayesian Ensemble, MCD Ensemble	Wisconsin Diagnostic Breast Cancer dataset (WDBC)	Bayesian Ensemble provided most reliable results	-	Limited data, use of transfer learning techniques
Leventi et al. [57]	Probabilistic neural network	Breast Cancer Wisconsin (Diagnostic) dataset	Accuracy: 95.32%	Test	Small dataset, low accuracy for breast cancer prediction
Hussain et al. [58]	Deep convolution networks	Mammography images	GoogleNet: 99.26% accuracy, AlexNet: 99.26% accuracy	Test	Small dataset, Binary prediction
Kayikci et al. [59]	Gated attentive multimodal deep learning	METABRIC dataset (1,980 records)	AUC: 0.95, accuracy: 91.2%, precision: 84.1%, recall: 79.8%	Test	Small dataset, Binary prediction

2.2.3. Breast cancer datasets

Table 2.2 includes a comparison between the utilized breast cancer prediction datasets, with information related to the studies used them and the best obtained result.

Table 2.2. A detailed comparison between utilized datasets.

Dataset	Specifications	Outcome (Target)	Studies Used	Best Result Obtained
BCSC	280660 records and 12 columns	Cancer prediction (Yes: 1, No: 0)	Shieh et al. [43], Rajendran [34] Li and Sundararajan [44], Kakileti et al. [32], Li et al. [13]	AUC increased to 0.65 (Shieh et al.) Accuracy: 99.1% with low sensitivity 91.1%
BCIMS	7,127 cases, 10 risk factors	Cancer prediction (Yes: 1, No: 0)	Hou et al. [33]	74.2% accuracy (XGBoost)
Geneva University Hospitals	112,587 individuals, 14 risk factors	Breast cancer lifetime risk predictions (near-population risk, Moderate risk, High risk)	Ming et al. [45]	Accuracy: 88.9%
Breast Cancer Wisconsin Diagnostic (BCWD)	899 records, 30 risk factors, 1 class	Cancer prediction (Yes: 1, No: 0)	Saleh et al. [48]	95.18% accuracy
Wisconsin Breast Cancer	569 records, 10 risk factors	Cancer prediction (Yes: 1, No: 0)	Uddin et al. [49] Hamedani et al. [56] Guo et al. [23] Leventi et al. [57]	98.79% accuracy
Special Dataset	5,176 records	Cancer prediction (Yes: 1, No: 0)	Botlagunta et al. [50]	83% accuracy, 0.87 AUC (DT classifier)
Haberman's survival	306 cases, 3 factors only	Survival status (1 or 2)	Al-Jawad et al. [18]	78.22% recall, 74.77% precision (BN)

Netherlands Cancer Registry	37,320 samples, 6 factors	Cancer risk prediction (Yes: 1, No: 0)	Annemieke et al. [55]	BNs outperformed regression in some cases
Cancer dataset	36,000 cases	Cancer risk prediction (Yes: 1, No: 0)	Yang et al. [52]	76.07% accuracy (Markov Model)
NCBI	1,580 samples (sequence data)	Cancer risk prediction (Yes: 1, No: 0)	Kurian et al. [53]	94.30% accuracy (DT)
BcBase and ORB	Not specified	Cancer risk prediction (Yes: 1, No: 0)	Savic et al. [54]	Centrally-trained models outperformed federated models
Simulated data, Newton-Wellesley Hospital, Cancer Genetics Network	Not specified	Genetic risk prediction (Yes or No)	Guan et al. [55]	Combination models outperformed individual models
Mammography images	Not specified	Cancer prediction (Yes, No)	Hussain et al. [58]	GoogleNet: 99.26% accuracy
METABRIC	1,980 records	Breast cancer classification (6 types luminal A, luminal B, HER2-enriched, basal-like and normal-like)	Kayikci et al. [59]	AUC: 0.95, accuracy: 91.2%

2.2.4. Related work summary

These studies introduced a broad range of predictive models for cancer risk, utilizing methods such as logistic regression, supervised machine learning algorithms, support vector machines, random forest, deep learning networks, XGBoost, and convolutional neural networks, among others. These previous studies were applied on several datasets of different sizes, from small ones with only a few patients to more extensive ones containing thousands of records.

The best models achieved accuracy rates ranging from the mid-70s to mid-90s percentile, with a few cases reaching above 99% accuracy and all these studies used a small dataset.

However, all these models focus on binary prediction, dividing cases into "cancer" or "non-cancer" categories.
The limitations of these approaches are evident in studies that highlighted low precision, low recall, and false positives or negatives as significant issues.
Additionally, while several datasets were utilized, many studies were limited by the small size of the datasets. Some other constraints highlighted included the limited number of risk factors and the binary nature of the predictions.
Based on these observations, a range-based prediction model for cancer risk could be a valuable addition to the current state of art in this field. A range-based model would be better suited to reflect the complex nature of cancer risk, which often isn't binary but exists on a spectrum. This model could potentially provide more accurate and reasonable information for decision-making and individualized patient care.

One of the primary focuses of the new model should be incorporating larger datasets to ensure robust training and validation.
Additionally, including more different risk factors, possibly integrating genomic, proteomic, and lifestyle data, could help create a more holistic risk prediction model.

As some studies noted, models that incorporated more risk factors tended to perform better, suggesting this could be a valuable idea to explore.
Furthermore, it would be interesting to analyze the impact of using a range-based prediction model on the precision, recall, and overall accuracy of predictions.
Therefore, I propose a research project aiming to develop and validate a range-based cancer prediction model.

This model should utilize a comprehensive set of risk factors and be trained on large, diverse datasets to provide a more refined and precise understanding of an individual's cancer risk. In doing so, it would address the limitations of previous studies and potentially offer a more effective tool for cancer prediction.

Chapter 3

Materials and methods

3.1. Dataset

The present study utilizes the BCSC dataset, which comprises 280,660 records and 12 risk factors detailed in Table 3.1. Additionally, the dataset contains a variable named "count," which indicates the frequency of each record within the dataset, as specified in the BCSC dataset. Table 3.1 shows the details of the BCSC datasets and its risk factors.

Table 3.1. Description of the breast cancer dataset.

No.	Risk Factor	Subcategory	Definition	Percentage
1	Menopause	Pre	0	23.58%
1	Menopause	Post or age>55	1	68.76%
1	Menopause	Unknown	9	7.66%
2	Age group	Group1	35-39	1.77%
2	Age group	Group2	40-44	12.1%
2	Age group	Group3	45-49	16.15%
2	Age group	Group4	50-54	17.9%
2	Age group	Group5	55-59	13.95%
2	Age group	Group6	60-64	11.1%
2	Age group	Group7	65-69	9.58%
2	Age group	Group8	70-74	8.48%
2	Age group	Group9	75-79	6.06%
2	Age group	Group10	80-84	2.91%
3	Density	Almost entirely fatty	1	6.19%
3	Density	Scattered fibro-glandular densities	2	32.69%
3	Density	Heterogeneously dense	3	28.17%
3	Density	Extremely dense	4	5.68%
3	Density	Unknown or other indexes	9	27.26%
4	Race	white	1	72.63%
4	Race	Asian/Pacific Islander	2	4.3%
4	Race	black	3	5.09%
4	Race	Native American	4	1.19%
4	Race	other/mixed	5	0.92%
4	Race	unknown	9	15.87%
5	Hispanic	No	0	73.1%
5	Hispanic	Yes	1	6.58%

5	Hispanic	Unknown	9	20.32%
6	BMI	10-24.99	1	21.27%
6	BMI	25-29.99	2	13.6%
6	BMI	30-34.99	3	6.05%
6	BMI	35 or more	4	3.25%
6	BMI	unknown	9	55.83%
7	Age at first birth (agefirst)	Age<30	0	30.18%
7	Age at first birth (agefirst)	Age 30 or greater	1	5.9%
7	Age at first birth (agefirst)	Nulliparous	2	8.41%
7	Age at first birth (agefirst)	unknown	9	55.51%
8	Number of first degree relatives with breast cancer (nrelbc)	zero	0	71.81%
8	Number of first degree relatives with breast cancer (nrelbc)	one	1	12.36%
8	Number of first degree relatives with breast cancer (nrelbc)	2 or more	2	0.65%
8	Number of first degree relatives with breast cancer (nrelbc)	unknown	9	15.18%
9	Previous breast procedure (brstproc)	no	0	71.97%
9	Previous breast procedure (brstproc)	yes	1	17.57%
9	Previous breast procedure (brstproc)	unknown	9	10.46%
10	Last mammogram before the index mammogram (lastmamm)	negative	0	75.22%
10	Last mammogram before the index mammogram (lastmamm)	false positive	1	1.42%
10	Last mammogram before the index mammogram (lastmamm)	unknown	9	23.36%
11	Surgical menopause	natural	0	30%
11	Surgical menopause	surgical	1	17.86%
11	Surgical menopause	unknown or not menopausal	9	52.14%
12	Hormone therapy	no	0	30.47%
12	Hormone therapy	yes	1	28.56%
12	Hormone therapy	unknown	9	40.97%
13	Count	-	Frequent of each record	

14	Cancer prediction	Cancer	1	3.32%
		No cancer	0	96.68%

The current study uses the entire BCSC dataset, taking into account the "count" column.

3.2. Methodologies

In the current study, I suggest three connected branches of the breast cancer prediction range-based model. The first branch of my study is a novel risk factor weighted-based breast cancer prediction model that is applied on the entire BCSC dataset. In the second branch, I suggest a novel approach to create a range-based machine learning breast cancer prediction model, while in the last step, I used the deep learning models to create a robust range-based cancer prediction model. Figure 3.1 shows the general architecture of the proposed study.

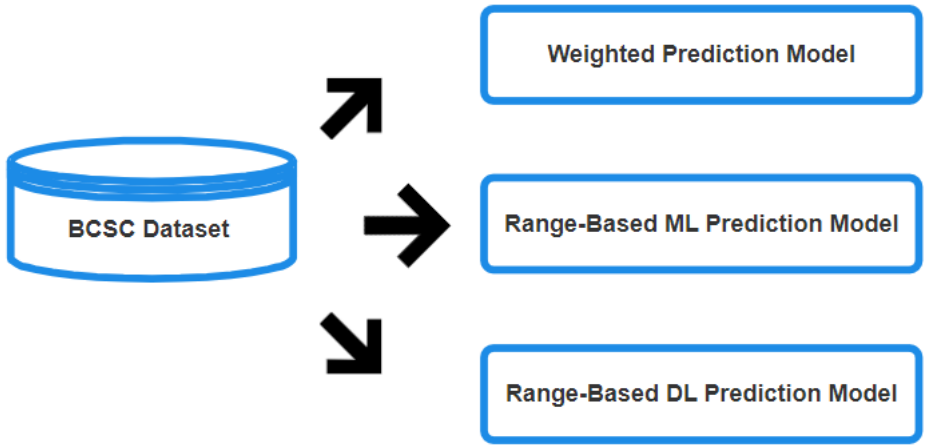


Figure 3.1. General branches of my study

3.3. First branch (Weighted-based breast cancer prediction model) proposed methodology

Figure 3.2 illustrates the risk-estimation model for breast cancer, where the BCSC dataset is sourced from http://www.bcsc-research.org/ and all risk factors are considered. To ensure that each risk factor has an equal impact on the final risk estimation, the dataset is normalized using Equation 3.1.

$$Risk\_factor_i = Risk\_factor_i / max(Risk\_factor_i); \quad i=1,2,\ldots,M \qquad (3.1)$$

Where M is the number of risk factors. The normalization step makes the value of each risk factor ranging from 0 to 1.

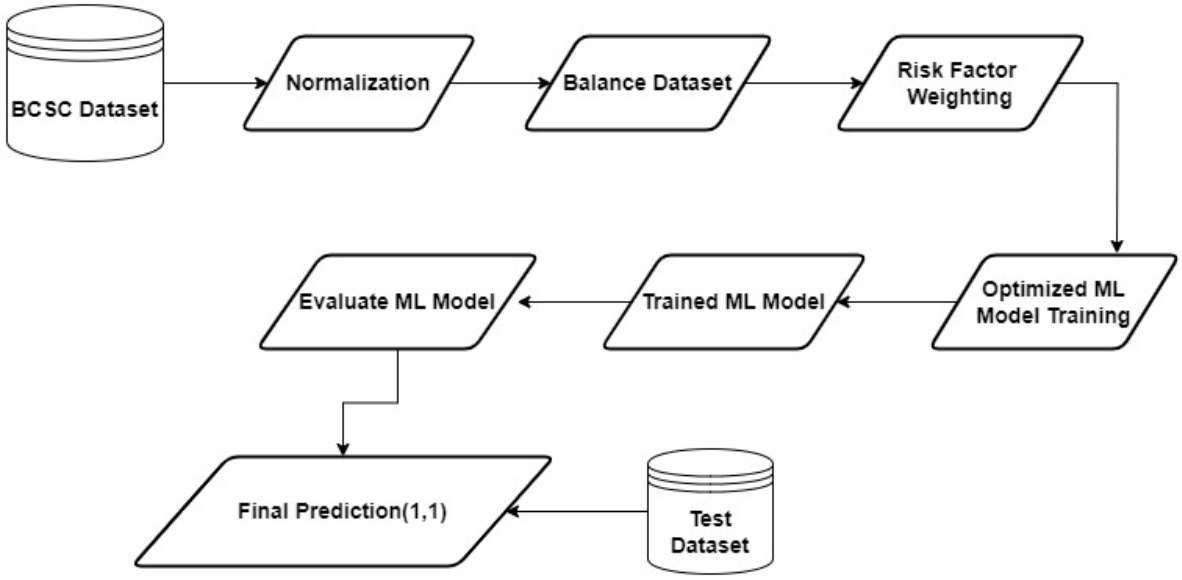


Figure 3.2. Weighted-based cancer prediction methodology.

The next step in the process involves balancing the dataset to address the significant imbalance between the two target categories in the original BCSC dataset. With only 3.32% of samples belonging to the "1" category and 96.68% belonging to the "0" category, the dataset is highly unbalanced, which can result in inaccurate predictions from any classifier trained on it.

To address this issue, three balancing approaches were used: oversampling, where the "1" category samples are duplicated many times to increase their percentage; down sampling, where some of the majority-class samples are removed to decrease their percentage; and a combination of oversampling and down sampling to achieve the desired balance. In this study, the oversampling technique involves selecting the minority class samples with the highest "count" value. These samples are considered to be more influential due to their higher representation within the minority class. By increasing their number through oversampling, the goal is to balance the class distribution and mitigate the potential bias caused by the class imbalance.

The third step involves applying a weighting algorithm to the dataset. Two materials were used to create an accurate weighting algorithm. First, a questionnaire was sent to 40 cancer specialists to establish medical knowledge and determine the impact of each risk factor on the final score of cancer risk. Second, international medical reports were analyzed to provide another perspective on the impact of breast cancer risk factors.

3.3.1. The proposed risk factor weighting methodology:

Equation 3.2 is used to define the degree of importance (DOI) of each risk factor based on the analysis of the questionnaire results. The DOI is calculated using the high-risk (Hi) and medium-risk (Mi) percentages of each risk factor.

$$DOI_i^Q = H_i * 0.6 + M_i * 0.4 \qquad (3.2)$$

The analysis of the questionnaire reveals that factors such as the number of first-degree relatives with breast cancer (nrelbc) and hormone therapy have the highest high-risk levels, while age, menopause, density, and race have the highest medium-risk levels.

Hispanic, breast procedure (brstproc), and surgical menopause are identified as having the lowest risk levels. Factors with a high DOI (more than 0.4) include nrelbc, age, and hormone therapy. Factors such as age at first birth, menopause, density, body mass index (BMI), last mammogram before the index mammogram (lastmamm), and race have medium DOI (between 0.3 and 0.4), while Hispanic, brstproc, and surgical menopause have low DOI (less than 0.3).

The international medical reports provide different opinions on the importance of risk factors. These reports were analyzed and the information about risk factors was compiled and classified according to the number of times the factors were mentioned in the list of essential risk factors (Ess_{Numi}) and secondary-risk factors (Sec_{Numi}).

The risk degree DOI_i^R was then calculated using Equation 3.3, which considers the essential and secondary risk factors. The value of DOI_i^R is calculated by summing up 90% of the essential risk factor effect and 10% of the secondary risk factors that have been identified in the medical studies that were analyzed.

$$DOI_i^R = \frac{1}{n} * (0.9 * \sum_{j=1}^{n} Ess_{Num_i} + 0.1 * \sum_{j=1}^{n} Sec_{Num_i}), \quad 1 \le j \le n \qquad (3.3)$$

Where n is the number of medical studies that have been analyzed.

The final DOI (DOI_i^F) is determined by combining the questionnaire-based degree of importance (DOI_i^Q) and the international medical reports-based degree of importance (DOI_i^R), as suggested in Equation 3.4. The suggested training weight (STW) is then inferred based on the final DOI (DOI_i^F), using Equation 3.5.

To summarize, the risk factors weights are determined based on the analysis of a questionnaire and international medical reports. The questionnaire-based degree of importance (DOI_i^Q) and the international medical reports-based degree of importance (DOI_i^R) are combined to calculate the final DOI (DOI_i^F), which is used to determine the suggested training weight (STW).

$$\mathrm{DOI}_i^F = \frac{DOI_i^Q + DOI_i^R}{2} \qquad (3.4)$$

$$STW_i = \begin{Bmatrix} \mathrm{round}(\mathrm{DOI}_i^F * \alpha) & if\ \mathrm{DOI}_i^F \geq \mathrm{T1} \\ \mathrm{round}(\mathrm{DOI}_i^F * \beta) & if\ \mathrm{DOI}_i^F \geq \mathrm{T2} \\ 1 & otherwise \end{Bmatrix} \qquad (3.5)$$

Where α and β are experimental values changing according to different datasets, different risk factors and different problems. For our problem $\alpha = 6$ and $\beta = 5$. T1, T2 are also experimental parameters with values 0.49 and 0.39, respectively. The selection of T1 and T2 parameters is done in an experimental way (i.e. these two values corresponds to the utilized BCSC dataset). T1 and T2 thresholds are also selected by means of observing the result of the DOI values of all risk factors in the dataset. So, the selection of such values is based on the DOI values (Refer to Table 4.1 to see the DOI values which are in range (0.165 - 0.4525) so the selection of good T1 and T2 result in a good weighting method. If we chose low T1 and high T2, it will produce unsuitable STW values.

3.3.2. Machine Learning Model

Once the final impact (weight) of each risk factor has been obtained, the final step is to select a machine learning (ML) model. Although there are multiple ML prediction algorithms available, the optimization tree model is chosen for its ability to tune hyperparameters, handle missing or noisy data, and manage redundant attribute values [60] [61]. The decision tree algorithm first considers all samples of the dataset as the root node. The decision tree algorithm starts by treating all samples of the dataset as the root node. The main challenge in this algorithm involves selecting the best attribute to serve as the root node and determining whether to split the node into all attributes and select the attribute with the best split performance. In order to choose the most suitable attribute, decision trees calculate the Information Gain (IG) across all possible attributes, as shown in Equation 3.6 [62], and select the attribute with the lowest IG value. This means that the selected attribute is the one that provides the best separation of the training samples.

$$IG(T,a) = H(T) - H(T|a) = -\sum_{i=1}^{k} p_i \, log_2(p_i) - \sum_{i=1}^{k} -p_r(i|a) \, log_2(p_r(i|a)) \qquad (3.6)$$

Equation 3.6 calculates the IG used in the decision tree algorithm. It involves several variables, such as H(T) which is the entropy of the parent node of the tree T, H(T|a) which is the entropy of the child node a (attribute a), k which is the number of subsets generated by each split, pi which is the percentage (probability) of class i in the node T, and pr(i|a) which is the percentage of class i given that the split child (attribute) is a.

In the case of the optimizable tree classifier, three parameters are tuned. These parameters include the criterion, which determines the attribute selection measure, the splitter, which determines the split strategy, and the maximum depth of a tree.

3.4. Second branch (Novel range-based breast cancer prediction model)

To improve the efficiency and accuracy of breast cancer prediction, we recommend using a range-based cancer score value instead of a binary scalar value (0 or 1). This approach would provide a range value between 0% and 100%, indicating the potential risk of breast cancer rather than a simple binary decision of either having cancer or not. The proposed methodology, which includes these suggestions, is illustrated in Figure 3.3.

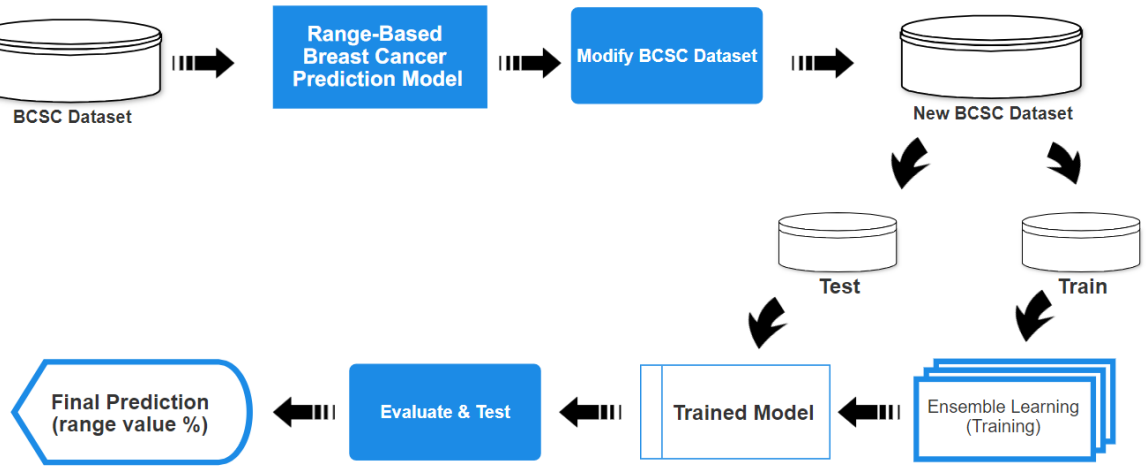


Figure 3.3. General steps of the proposed Ensemble range-based prediction model

3.4.1. Design the breast cancer range-based model

Building upon our defined risk factor weighting system, we introduce a range-based prediction model. This model aims to capture the multifaceted nature of breast cancer risk by incorporating the weighted importance of individual risk factors into a composite score, as detailed in the subsequent equations and methodology.

The overarching goal of this range-based score is to offer a more detailed understanding of breast cancer risk, hence the introduction of a score-based prediction system. The methodology is built upon two main systems.

Firstly, the breast cancer factors weighting system, derived from the prior section (3.2.), provides insight into how different factors play a role in breast cancer risk.

Secondly, the statistical system aims to combine these weighted factors to compute a composite breast cancer risk score (range-based one), encapsulating multiple aspects of the risk profile. Figure 3.4 illustrates this range-based cancer prediction model.

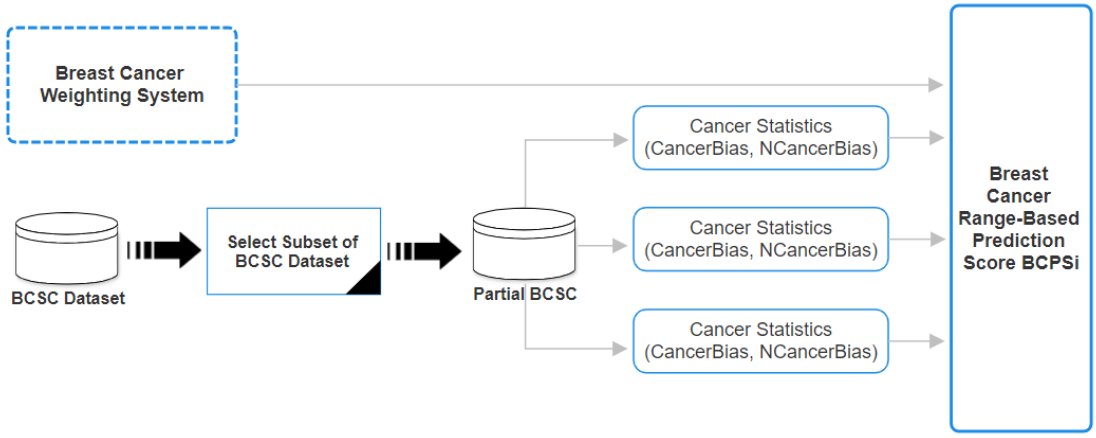


Figure 3.4. Range-based cancer prediction methodology

This approach reinforces the point that the scoring system is not just a mathematical construct but has its foundations in real-world perceptions (via the questionnaire) and established medical knowledge (via international reports). It emphasizes that the methodology is rigorous, comprehensive, and designed with scientific soundness in mind.

To ensure that the risk score obtained is correct and accurate, a selected subset of the entire dataset is used. This approach ensures that the "cancer" class has a higher percentage than the "non-cancer" class, thus providing sufficient information to the proposed probability model and ensuring that the final range-based cancer score relies primarily on this information. The selection of this subset from the BCSC dataset is based on two principles. The first principle involves selecting all samples from the "cancer" class. The second principle involves selecting the "non-cancer" samples with the highest values of the "count" attribute, which indicates how frequently each sample is repeated in the entire dataset.

As a result of this step, we get a subset obtaining 67633 samples with 68.88% as the "cancer" class percentage.

The inputs for our breast cancer prediction system include the scalar weights assigned to each risk factor obtained from previous model (First branch weighted-based breast cancer prediction model). These weights indicate the level of importance of each risk factor and will be used in the prediction system to ensure an accurate prediction score.

Additionally, we consider the general probability of cancer and non-cancer, denoted by $cancer_{Bias}$ and $Ncancer_{Bias}$, respectively.

These probabilities are based on the breast cancer dataset and reflect the previous knowledge obtained from the BCSC sub dataset.

The likelihood of each risk factor given the prediction result of cancer or non-cancer is also taken into account. This likelihood is calculated by summing all the inner values of the risk factor's probability, as shown in Equation (3.7), where k represents the total number of inner values for the risk factor.

$$P(Risk\_Factor_i \mid Prediction) = \Sigma_k P(Inner\_Value_{ij} \mid Prediction) \quad (3.7)$$

The probability of each risk factor's effect on the final breast cancer score is based on medical opinion and is denoted as $P(pre\_cancer_{ij})$. This information is obtained from the analysis of medical questionnaires administered to specialist physicians in breast cancer. The final breast cancer prediction score is then calculated as a range-value using the inputs described above, as shown in Equation (3.8).

$$BCPS_i = cancerBias * BCPS_{cancer} + NcancerBias * BCPS_{Ncancer} \quad (3.8)$$

According to Bayes' theorem, the post probabilities of cancer and non-cancer, denoted as BCPScancer and BCPSNcancer, respectively, are calculated based on the risk factors using Equations 3.9 and 3.10.

$$BCPS_{cancer} = \Sigma_n P(prediction = cancer \mid Risk\_factor_i) \times (STW(j) / \Sigma_n STW(j)) \quad (3.9)$$

$$BCPS_{Ncancer} = \Sigma_n P(prediction = Non\text{-}cancer \mid Risk\_factor_i) \times (STW(j) / \Sigma_n STW(j)) \quad (3.10)$$

The recommended training weight, STWj, from our previous weighted-based model is used in conjunction with the total number of risk factors, denoted by n, to calculate the post probability of each risk factor. This calculation is performed using Equation 3.11.

$$P(Prediction = Cancer \mid Inner\_value_i) = \Sigma_k (P(Inner\_value_{ij} \mid Prediction = Cancer) \times P(Pre\_cancer_{ij}) / P(Inner\_value_{ij})) \quad (3.11)$$

The number of inner values for a risk factor, denoted by K, determines the number of probabilities to be calculated. For example, the menopause risk factor has three inner values (K=3): Pre-menopause (0), Post-Menopause (1), and Unknown (9).

The pre-probabilities of cancer related to each inner value, denoted as $P(Pre\_cancer_{ij})$, are based on previous knowledge. The evidence of each risk factor's information, denoted as $P(Innervalue_{ij})$, is calculated using Equation (3.12).

$$P(Inner\_value_{ij}) = P(Inner\_Value_{ij} \mid Prediction = Cancer) \times P(Pre\_Cancer_{ij}) + P(Inner\_Value_{ij} \mid Prediction = Non\text{-}Cancer) \times (1 - P(Pre\_Cancer_{ij})) \quad (3.12)$$

3.4.2. The new BCSC version

In this step, three new attributes are added to the BCSC dataset, namely the cancer score, non-cancer score, and final prediction. These additions enhance the dataset and provide valuable information for future studies to predict and analyze the BCSC dataset. The final prediction of our proposed methodology will utilize this updated version of the BCSC dataset.

3.5. Train the ensemble learning model using the new ranged dataset

Ensemble learning is a powerful method that combines multiple classifiers or models to improve performance. This approach has gained attention in recent years, particularly when combined with hyperparameters optimization [63]. Several hyperparameters are selected for optimization, including the maximum number of splits, number of learners, and learning rate. The ensemble method used in this study is the AdaBoost algorithm [64], with decision trees as the learner type [65].

The Decision Tree (DT) model is a type of machine learning model in which features are represented by internal nodes, and branching represents potential outcomes [66].

At the beginning of the process, the most promising feature is selected as the root node, and the splitting process is applied using a specific criterion.

Multiple learners are created and learned sequentially by fitting the model to the dataset [67] [68].

In each step, a decision tree learner is selected and fitted to minimize errors, and the resulting misclassified samples are used to train the next learner. This approach ensures that misclassified samples from previous models are correctly classified by subsequent ones [67].

3.5.1. Bagged and Boosted Trees [69]

As mentioned in section 3.5, ensemble methods involve combining multiple decision trees to achieve better predictive performance than using a single decision tree. The fundamental principle behind the ensemble model is that a group of weak learners can work together to form a strong learner.

There are two main types of ensemble learning; bagging and boasting.

Bagging (or Bootstrap Aggregation) is a technique used to reduce the variance of a decision tree. The idea behind bagging is to create multiple subsets of data from the training sample, chosen randomly with replacement. Each subset of data is then used to train a separate decision tree, resulting in an ensemble of different models. The average prediction from all the trees is used, making it more robust than a single decision tree.

Random Forest is an extension of bagging that adds an extra step. In addition to taking a random subset of data, it also randomly selects features to grow trees, rather than using all features. When many random trees are grown, it is called a Random Forest.

Boosting is an ensemble technique that involves creating a set of predictors. With this technique, learners are learned sequentially, with early learners fitting simple models to the data and analyzing the data for errors. In other words, consecutive trees are fitted to the data (using a random sample), and at each step, the goal is to address the model error from the previous tree.

If an input is misclassified by a hypothesis, its weight is increased so that the next hypothesis is more likely to classify it correctly. By combining the entire set of hypotheses at the end, weak learners are converted into a better-performing model.

Figure 3.5 shows the difference between bagging and boasting decision trees ensemble.

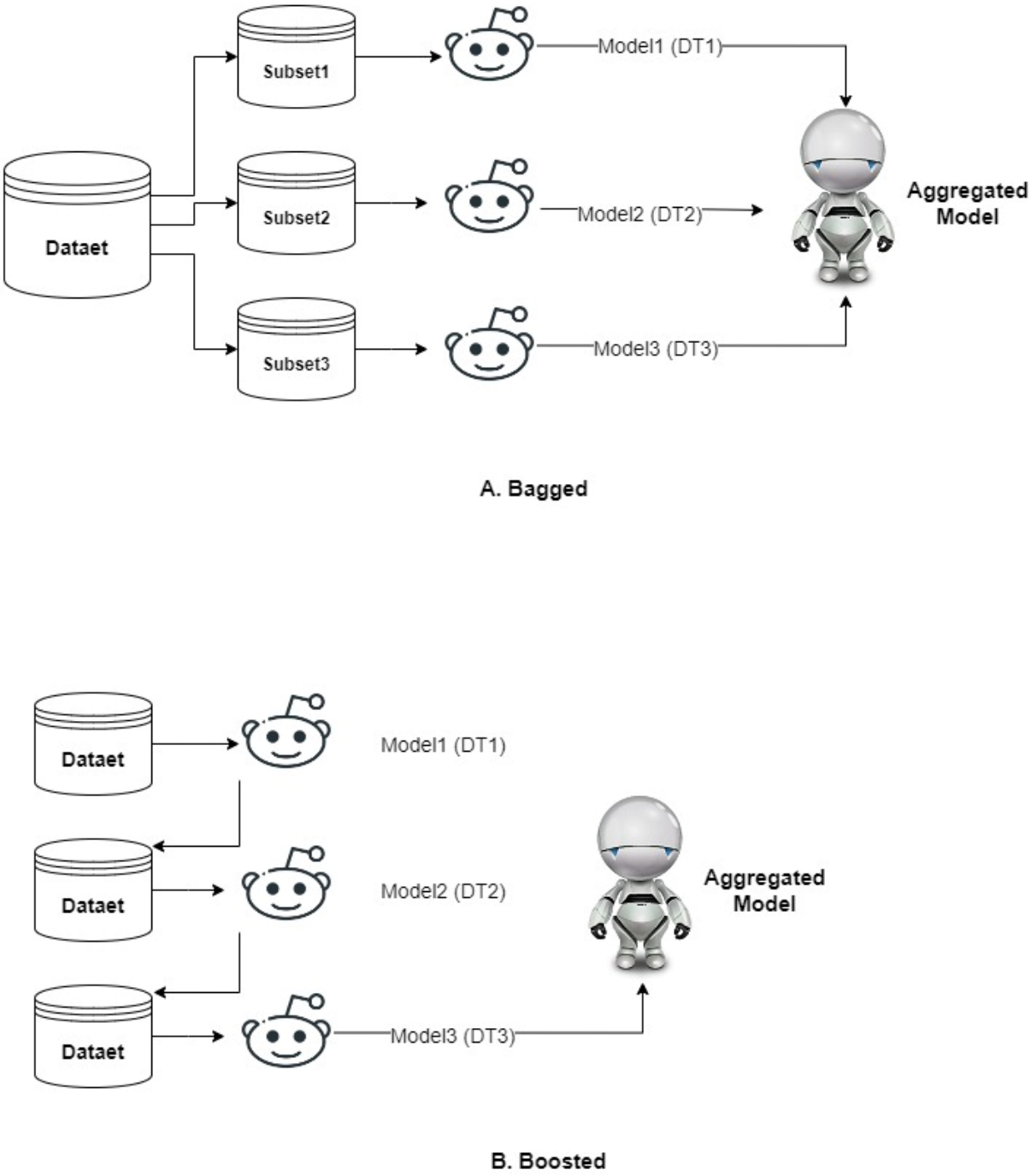


Figure 3.5. Bagged and boosted decision trees ensemble

3.6. Third branch (Range-based deep learning model)

The proposed breast cancer range-based deep model consists of five main steps, as depicted in Figure 3.6. The first step involves dataset preprocessing, which involves grouping the target or

classes into specific categories. In the second step, the dataset is divided into training and test sets. The third step involves constructing and training the Deep Learning (DL) architecture using the training set. In the next step, the ensemble Machine Learning (ML) model is built and trained using the same training set. In the final step, the scores from the DL and ML models are combined using score-level fusion, and the final prediction is computed.

3.6.1. Preprocessing

In this step, the BCSC dataset obtained from section (3.4). The output of the second branch of my study is the modified and balanced version of the BCSC dataset (in which target column is a value of different multiple range-based scores instead of 0/1 categorization). Besides, the dataset is balanced from the output of the first branch.

To simplify the classification problem, the target column of the BCSC dataset is grouped into categories, reducing the number of range-based categories to simplify the calculations. After preprocessing the dataset, it is split into two sub-datasets, with a 20% percentage for the test set and 80% for the training set. In case of using LSTM model (Deep learning model), another subset (validation set) is also considered so the split became: 60% as a training set, 20% as a validation set and 20% as a test set.

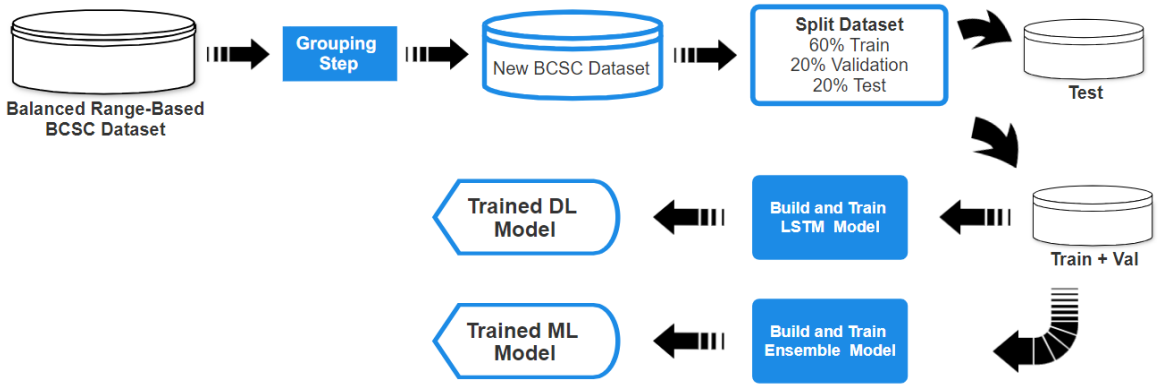


Figure 3.6. Proposed breast cancer range-based deep prediction model

3.6.2. LSTM deep learning model

In this step, a specific Deep Learning (DL) architecture is proposed, which is illustrated in Figure 3.7. The first layer of the model is the sequence input layer, which takes the input features of the training samples and passes them to the next Long Short-Term Memory (LSTM) layer.

The LSTM layer is the main component of the DL model and consists of 500 neurons, each containing four basic cells: input cell, memory cell, forget cell, and output cell. The input cell receives input from the previous cell, while the forget cell determines what information to keep and what to discard, controlling the cell state reset.

Based on information received from the input cell, the LSTM model uses the forget cell, and previous hidden LSTM cell h_{t-1} to decide what to pass and what to forget. Memory cell collects information from previous time steps and helps to maintain the context. It updates its state based on both input and forget cells.

The output of the LSTM cell at a specific time step t is represented by c_t (output at time t). Additionally, another output is produced, represented by h_t, which is the current hidden LSTM output that will be passed to the next LSTM layer [70].

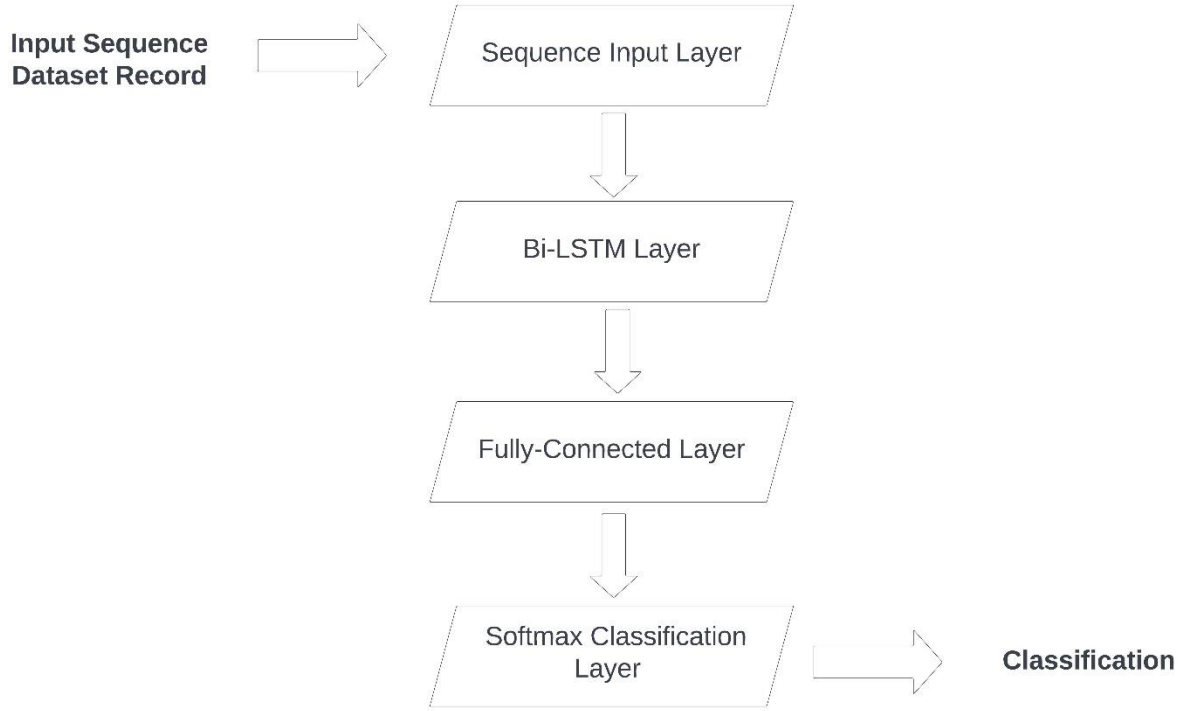


Figure 3.7. Proposed DL model

3.6.3. Bi-LSTM

A Bidirectional Long Short-Term Memory (Bi-LSTM) model is a type of Recurrent Neural Network (RNN) architecture that is actually developed from LSTM. Bi-LSTM was mainly used for processing sequential data. It consists of two LSTM layers, one of them reads the input sequence from left to right (forward LSTM), while the other one reads the input sequence from

right to left (backward LSTM). The output from each LSTM layer is concatenated to produce the final output of the Bi-LSTM layer [71].

The forward LSTM layer processes the input sequence in the forward direction, starting with the first element and processing each subsequent element in order. On the other hand, the backward LSTM layer, processes the input sequence in reverse order, starting with the last element and processing each preceding element in reverse order.

By processing the input sequence in both directions, the Bi-LSTM model can acquire both past and future context information, making it a powerful model for tasks such as natural language processing, speech recognition, and time series prediction.

Bi-LSTM is particularly useful for tasks where the context of a word or sequence is important to the meaning of the complete input sequence.

3.6.4. Ensemble ML model

Ensemble models are a type of Machine Learning (ML) model that utilizes a combination of multiple ML classifiers to produce a single classification. This methodology operates in two different ways to arrive at a final classification: boosting and bagging. In the boosting approach, the classification decision is based on an iterative strategy, where the first classifier introduces its decision to the next one, which learns from the first classifier's error and tries to minimize the classification error until the final classifier produces the final decision with the minimum classification error. In the bagging approach, the classifiers work in parallel, and the ensemble attempts to minimize prediction variance by generating new samples of the training dataset by repeating the training data and producing sub-datasets to train multiple classifiers. The final decision is based on the fusion of their scores.

In my study, I propose using an ensemble of 30 boosting decision trees, and to achieve optimal performance, we suggest using hyperparameters optimization for those 30 decision tree models.

3.6.5. Fusion model

Once the Machine Learning (ML) and Deep Learning (DL) models are constructed, they are combined using score-level fusion, where the ML and DL scores are merged to make the final prediction decision. Fusion is illustrated in Figure 3.8. The final score represents the weighted sum of the ML score and DL score as described in Equation 3.13.

$$Fusion\_score = W_1 * S_1 + W_2 * S_2 \qquad (3.13)$$

Where; W1, W2 are the weights of the ML and DL models, while S1 and S2 are the scores obtained by the evaluation process of the ML and DL models, respectively.

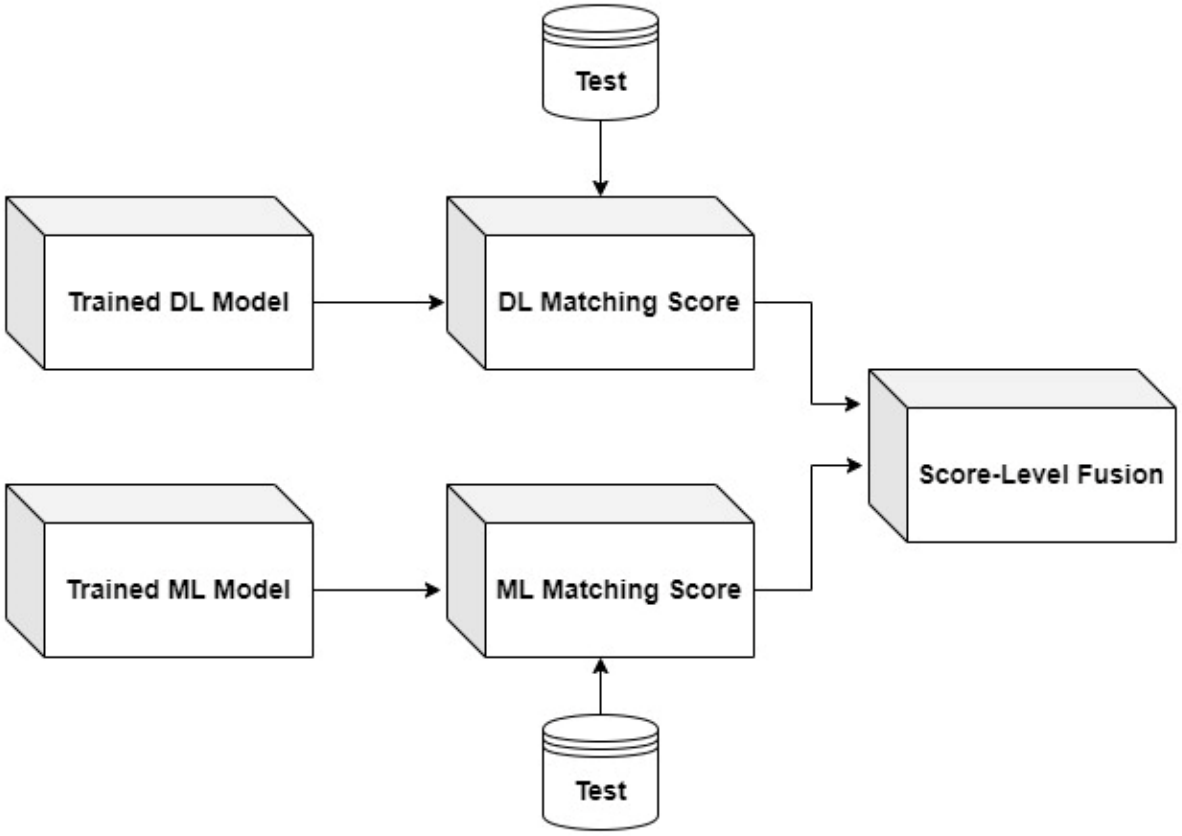


Figure 3.8. Proposed ML-DL fusion model

3.7. Fourth branch (Classification-based range-based ensemble model on the original dataset)

In this branch, the original dataset will be utilized again, and the probabilistic model will be used and applied to the dataset in order to compute the new distribution of the target column but without any balancing operations before this.
After that, the new modified dataset will be split into training (80%) and test (20%) sets. The training set will be balanced using oversampling approach (SMOTE algorithm).

In the third step, three different ML models will be trained using the training dataset. An ensemble model of these three ML models will also be created. 1D-CNN and LSTM DL models will also be trained using the training dataset. Similarly, an ensemble model of the two trained DL models will be built. All trained models will be evaluated using the performance metrics: accuracy, precision, recall and F1-scrore. Besides that, the Violin, the variance, the test score distribution and the distribution of the predicted and the actual breast cancer score will be all used to evaluate the trained models. Figure 3-9 shows the proposed methodology of the fourth part of the study.

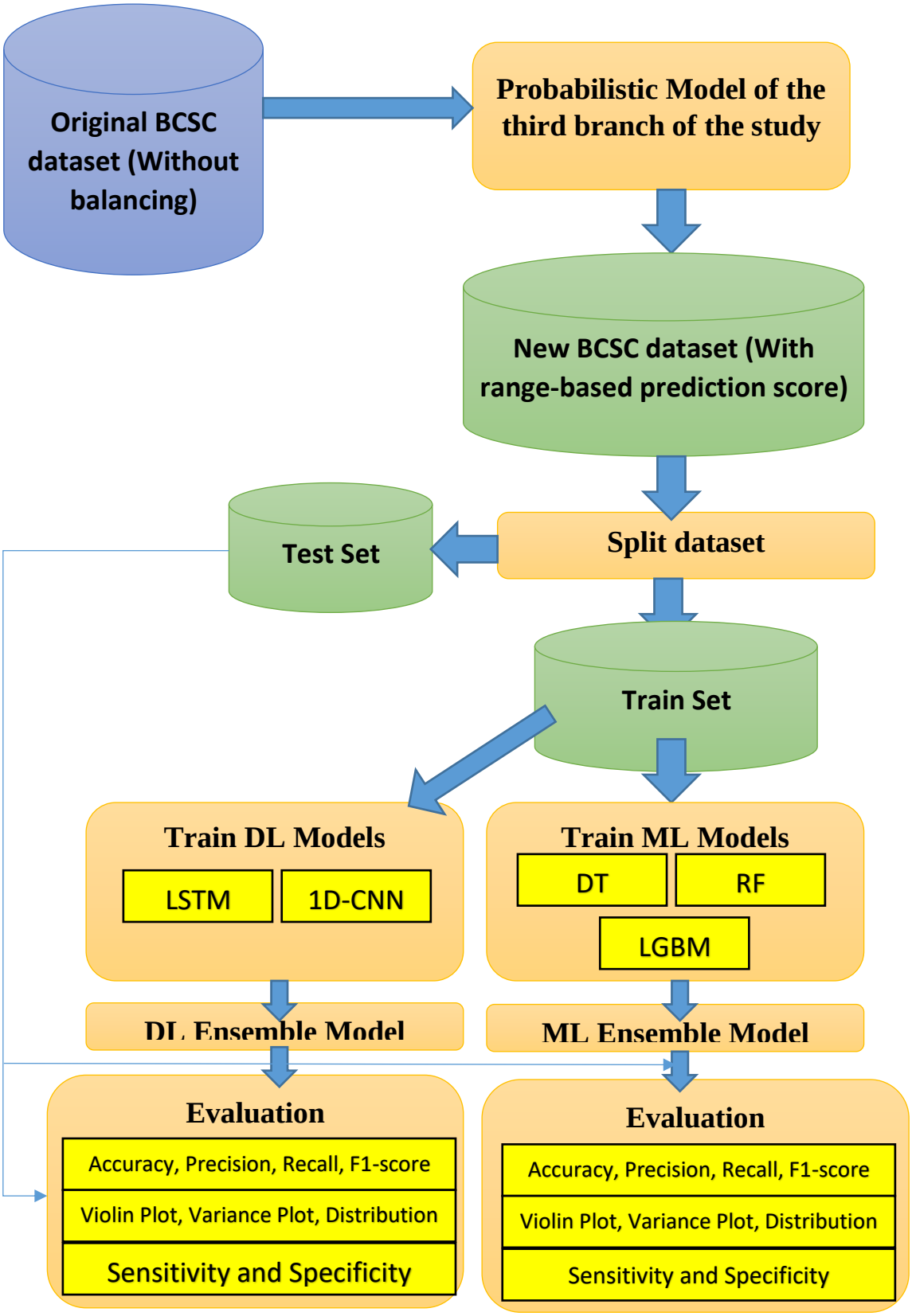


Figure 3.9. The Proposed methodology of the fourth and fifth part of the study

3.8. Fifth branch (Regression-based range-based ensemble model on the original dataset)

In the fifth part, the same dataset of the fourth part is used. The only difference here is that we are performing regression task so the target column will preserve its values without merging the adjacent categories.

Three regression models are proposed (Decision trees regression DTR, Random Forest regression RFR, and the K-NN regression models).

An ensemble of the three models will also be built.

All models will be trained using 80% of the dataset, while the rest (20%) will be used in the evaluation process a test set.

The distribution of the actual and predicted breast cancer scores will also be used to assess the performance.

3.9. Performance evaluation method

In order to evaluate the proposed ML and DL models, I computed many performance metrics. Here are the used metrics:

Several metrics, including True Positive Rate (TPR), False Negative Rate (FNR), Positive Predictive Rate (PPR), and False Discovery Rate (FDR) [72] [73] , are used to assess the performance of the model. These metrics are calculated using four statistics:

TP (true positives), which describes the correctly classified samples of the whole positive ones.

FN (false negatives), which represent the incorrectly classified samples of the whole positive ones.

TN (true negatives), which calculate the correctly rejected samples of the whole negative ones; and FP (false positives), which signify the incorrectly accepted samples of the whole negative ones.

TPR (TP/(TP+FN)) represents the proportion of correctly classified samples per predictive class, while FNR (FN/(TP+FN)) represents the proportion of incorrectly classified samples per true class.

Similarly, PPR (TP/(TP+FP)) is the proportion of correctly classified samples per predictive class, while FDR (FP/(TP+FP)) is the proportion of incorrectly classified samples per predictive

class. Additionally, accuracy is calculated as ((TP+TN)/(TP+TN+FP+FN)), representing the proportion of correctly classified samples out of all data samples [74].

Area Under Curve (AUC) [75]: AUC stands for Area Under the ROC Curve, which is a graphical plot of the model's performance in distinguishing between positive and negative classes. The ROC (Receiver Operating Characteristic) curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for all test samples. AUC represents the overall performance of the model across all possible test samples. AUC ranges from 0 to 1, with a higher value indicating better performance. It can be interpreted as the probability that a randomly chosen positive example will be ranked higher than a randomly chosen negative example by the model.

Confusion Matrix [76]: confusion matrix is a table that concludes the performance of a classification model by comparing the predicted targets with the actual targets (labels). It contains four values: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP and TN represent the number of correctly classified examples, while FP and FN represent the number of incorrectly classified examples. The confusion matrix is useful for calculating other metrics such as accuracy, precision, recall, and F1-score.

Regression metrics: in this study, two regression metrics are used; the mean squared error (MSE) and the median absolute error (MedAE).

Statistical analysis using violin plots, variance plots, distribution of the predicted and actual scores, besides the sensitivity and specificity analysis.

3.10. Utilized Software and Hardware

I used the following software and hardware in the current study:

CPU (intel core i5 4200U CPU @ 1.60GHz, 8 GB of RAM), Matlab 2020a, including the machine learning and deep learning toolbox, Specific medical Questionnaires.

Google Colab with python programming language.

For deep learning: GPU (NVIDIA GeForce 750 M) is used.

Chapter 4

Results and discussion

4.1. Introduction

In this chapter, all my previous mentioned methodologies will be experimented and evaluated. Three different branches are proposed so, there will be three different branches of results will be conducted. First, I will introduce the results of weighted-based breast cancer prediction system. Then, the results of the range-based prediction model will also be introduced and finally, the results of ensemble and deep learning model fusion will be listed. All results will be discussed.

4.2. Results of the weighted-based breast cancer prediction methodology

4.2.1. Results of Balancing BCSC dataset

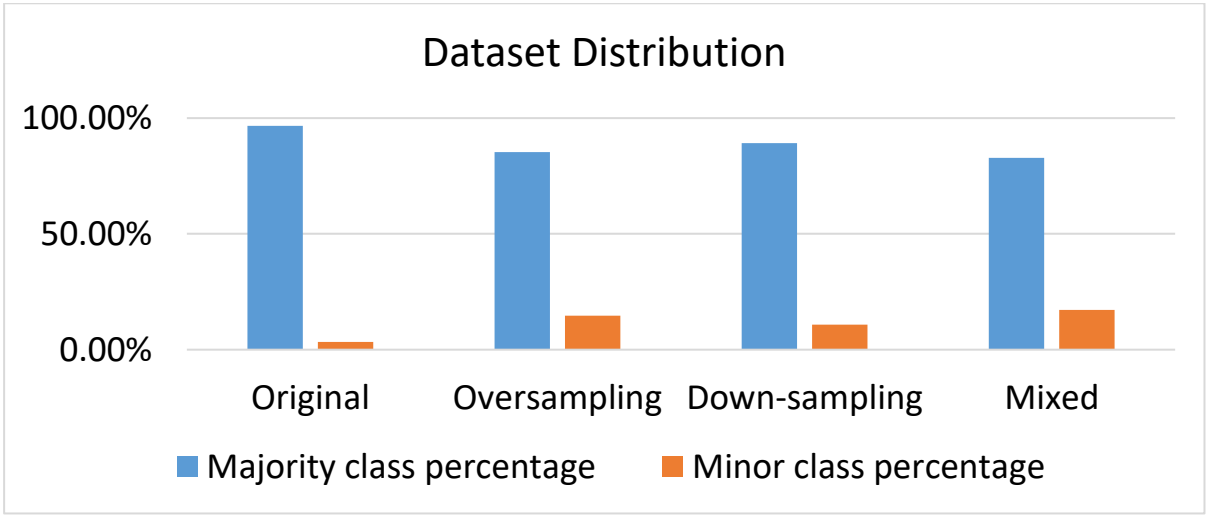
There are two fundamental steps involved in preprocessing the risk estimation dataset, namely normalization and balancing. The results of the suggested balancing methods are presented in Figure 4.1, with the majority class label being 0 (no cancer) and the minor class label being 1 (cancer risk).

To implement the oversampling approach, the "1" minor class was replicated **five times** until its percentage reached 14.64%, while the majority-class percentage became 85.36%. In contrast, the down-sampling approach involved reducing the majority-class samples by **a factor of 3.524** until the minor class reached a percentage of 10.78%, and the majority class reached 89.22%.

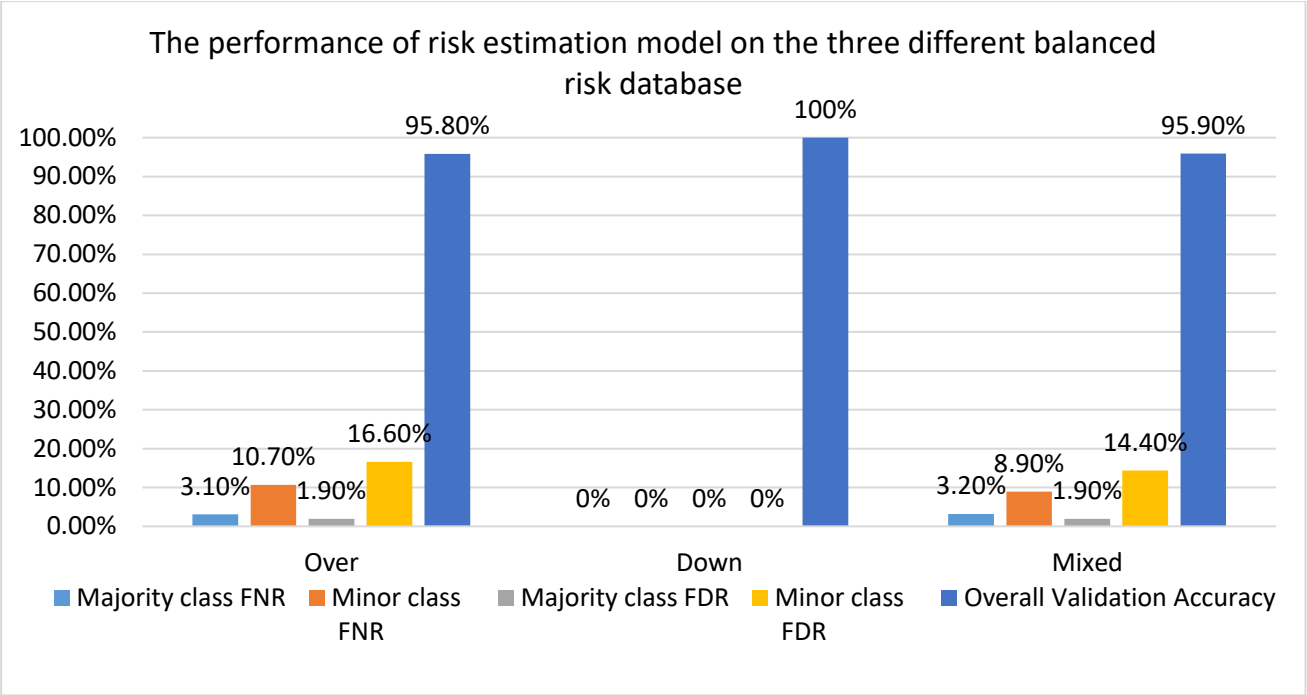
The last approach involved duplicating the minor class samples and removing some of the majority class samples until the minor and majority classes reached 17.1% and 82.9%, respectively. The number of majority class samples after balancing is 271355, 77000 and 225562 for oversampling, down sampling, and mixed cases, respectively. While for the minority class, the number of samples after balancing is 46525, 9305 and 46525 for the same sampling methods.

The utilized BCSC dataset is extremely unbalanced since the minor class constitute more than 95% of the samples which can bias the learning process to always predicting it as the target. Now, if we used the default balancing approach in which the minor class is oversampled until its percentage became 50%, we need to add too much generated records to satisfy this requirement (this may lead to data leakage or producing too much noise or repeated rows). So we chose to increase the minor class until it almost reaches 15% of the entire dataset. However, we can't get exactly 15% since the utilized oversampling approach was based on increasing the number of samples not define specific percentage. For mixed case, we removed samples of the major class and increased the number of minor class's samples. We used the same number of new added records of the minor class (46525 samples) but by decreasing the major class samples by a factor of almost 3.5, we get these percentages of major and minor classes.

Figure 4.1-B includes a detailed comparison of the performance of scaling choice (age=2, race=4, Hispanic=5, bmi=2, agefirst=3, nrelbc=3, Current_hor=2, menopause=0.5, Density=0.3, brstproc=0.2, lastmamm=0.3, surgmeno=0.2) over the three balanced datasets. Figure 3 indicates that the down-sampled dataset has the highest accuracy (100%) and the least error rates (0%); however, this down-sampled dataset has a volume of 27.15% only compared with the over-sampled version. So, although the down-sampled dataset has the best accuracy, the over-sampled and the mixed versions have better performance since they consist of a much larger number of samples so that the new test samples will be classified more correctly.



A.



B.

Figure 4.1. BCSC dataset distribution and performance measures comparison before and after balancing: A. Distribution, B. Performance measure

4.2.2. Weighting system results

The risk factors questionnaires degree of importance (DOI^{q}), medical reports degree of importance (DOI^{R}), and final mixed degree of importance (DOI^{f}) obtained from the weighting methodology are shown in Table 4.1.

Table 4.1. Results of weighting system (DOI^{R}, DOI^{F}, DOI^{Q}, STW).

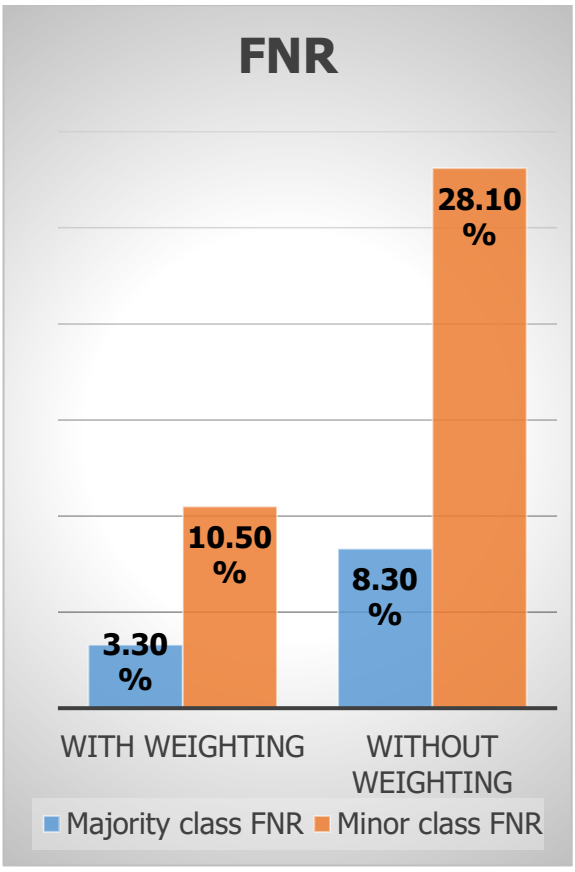
No.	Risk Factor	Medical records-based DOI									Questionnaires-based DOI				DOI^{f} **	STW
		Essential*				Secondary*				DOI^{R}	H	M	L	DOI^{Q}		
		1	2	3	4	1	2	3	4							
1	**Menopause**			1		1	1	1		0.3	30	47.5	22.5	0.37	0.335[7]	1
2	**Age group**	1	1	1	1					0.9	27.5	62.5	10	0.415	0.6575[1]	4
3	**Density**	1	1					1	1	0.5	25	45	30	0.33	0.415[6]	1
4	**Race**	1	1		1					0.675	25	40	35	0.31	0.4925[3]	3

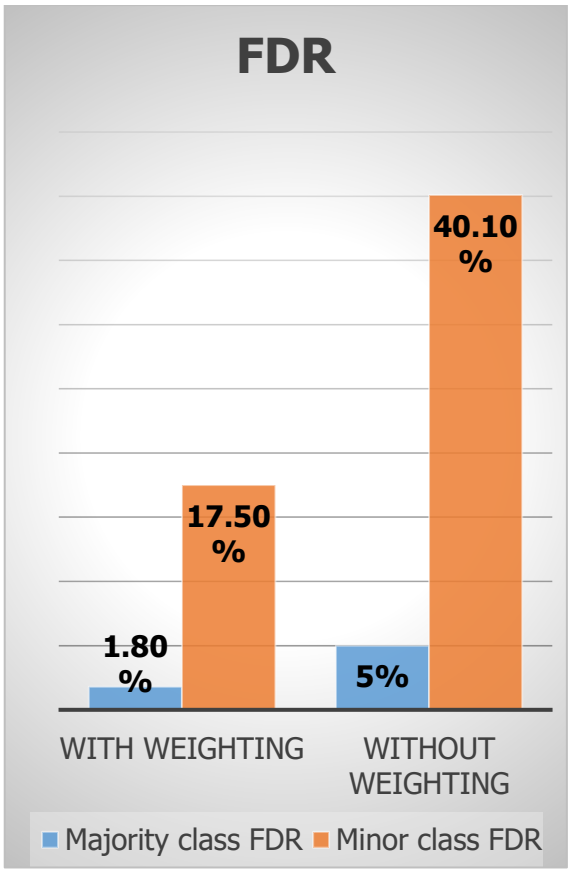
5	**Hispanic**	1						1	0.25	19.4	16.7	63.9	0.183	0.2165[9]	1
6	**BMI**		1			1		1	1	0.3	25.6	38.5	35.9	0.307	0.3035[8]	1
7	**agefirst**		1	1		1			1	0.5	27.5	45	27.5	0.345	0.4225[5]	2
8	**nrelbc**		1	1	1	1				0.7	56.4	25.6	17.9	0.44	0.57[2]	3
9	**brstproc**					1		1		0.05	34.2	23.7	42.1	0.3	0.175[10]	1
10	**lastmamm**									-	34.2	32.1	33.7	0.33	0.165[11]	1
11	**Surgical menopause**					1				0.025	7.7	30.8	61.5	0.169	0.097[12]	1
12	**Hormone therapy**	1	1					1	1	0.5	42.5	37.5	20	0.405	0.4525[4]	3

*Each 1 value indicates that this risk factor is assumed as essential or secondary factor in a study., **Numbers 1-12 in DOI^F indicates the weight order. H, M, L represents the High, Medium, and Low of DOI^R.

Table 4.2 shows the most significant risk factors for breast cancer, which include Age group, nrelbc, and race, while the risk factors of medium significance are Hormone therapy, agefirst, density, Menopause, and BM. Conversely, the least essential risk factors are Hispanic, brstproc, lastmamm, and surgical menopause.

Figure 4.2 illustrates the effects of weighting the risk factors against the non-weighted version of the dataset. The results indicate that the performance improves by 6.9% with the weighting approach. Additionally, the False Discovery Rate (FDR) is reduced by 22.6% and 3.2% for the minor and majority classes, respectively. Furthermore, the False Negative Rate (FNR) is minimized for both the majority and minor classes by 5% and 17.6%, respectively.


FNR
28.10 %
10.50 %
8.30 %
3.30 %
WITH WEIGHTING
WITHOUT WEIGHTING
Majority class FNR
Minor class FNR


FDR
40.10 %
17.50 %
1.80 %
5%
WITH WEIGHTING
WITHOUT WEIGHTING
Majority class FDR
Minor class FDR

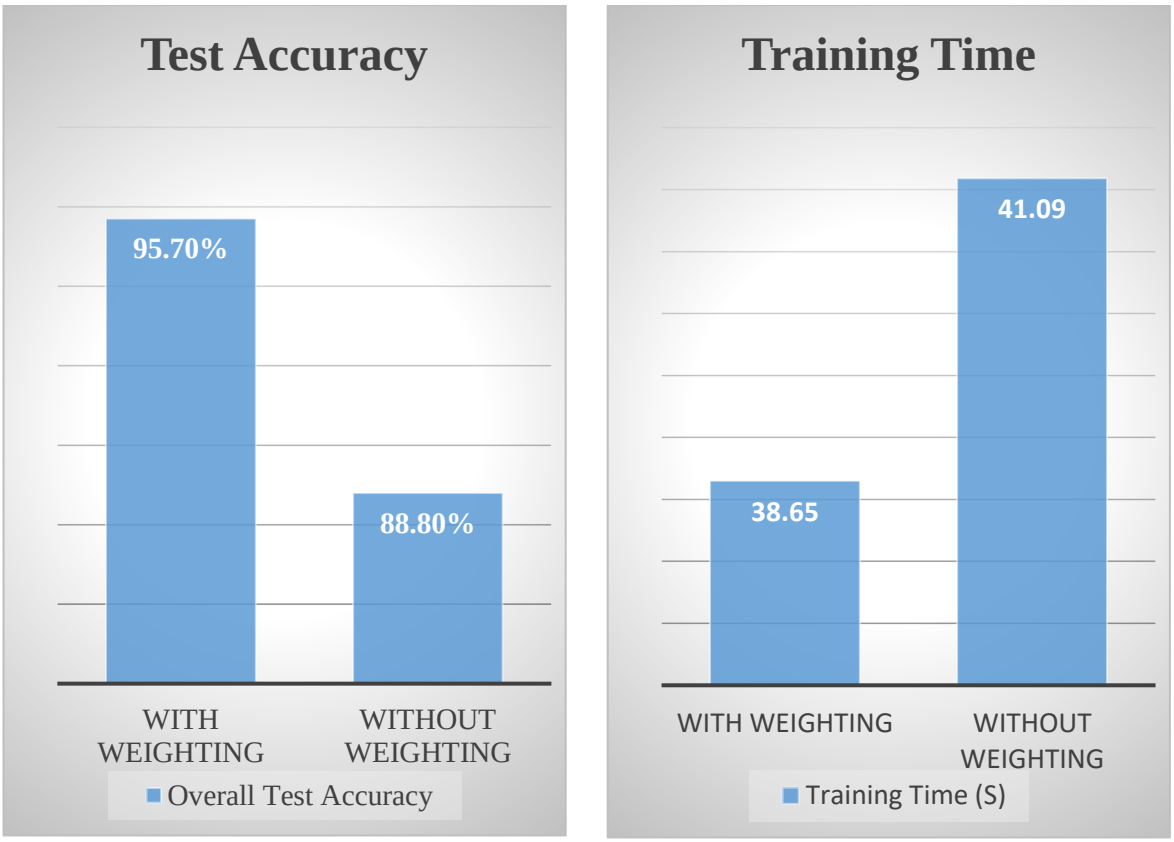


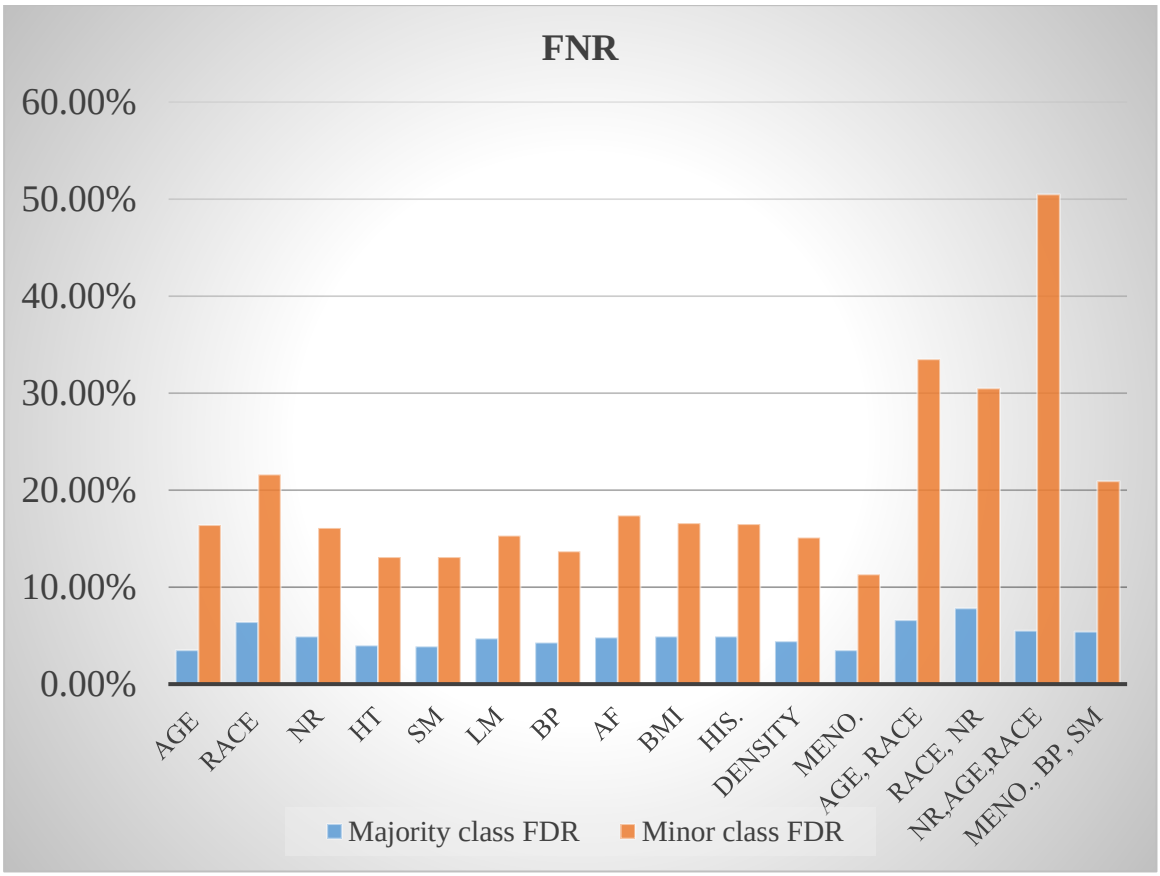
Figure 4.2. Results of Weighting-based breast cancer prediction model

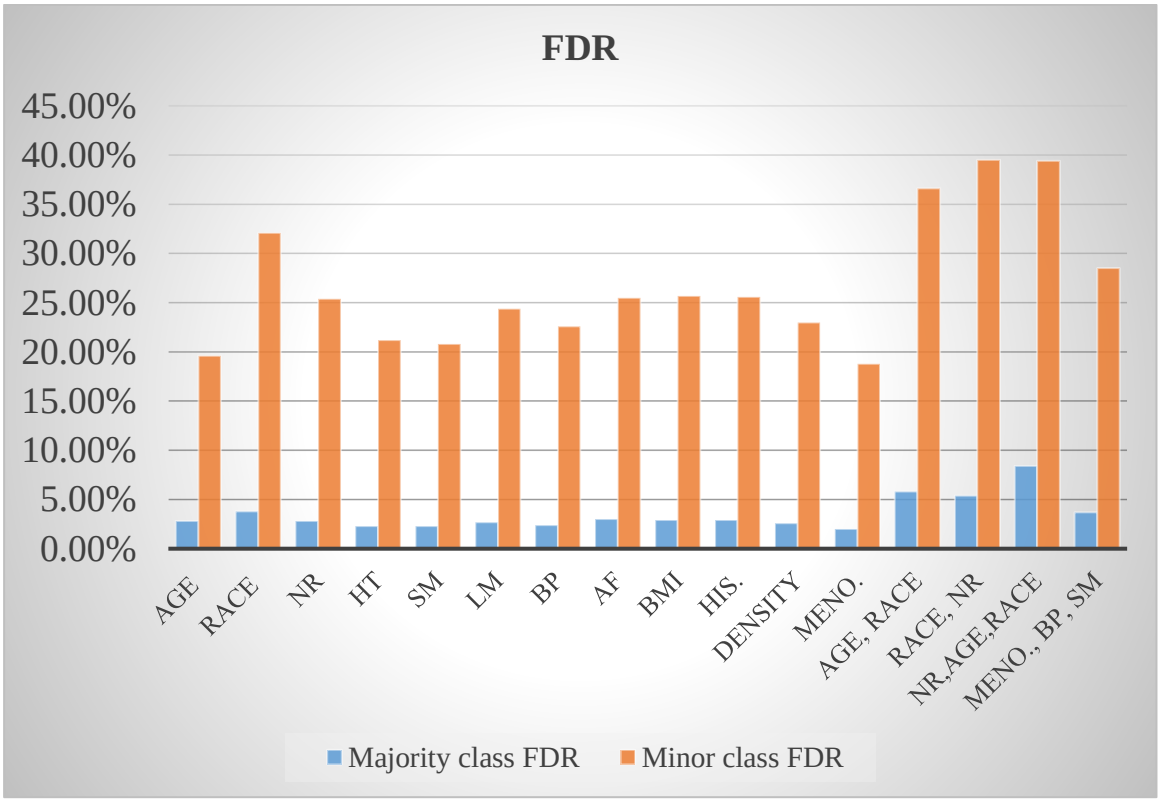
4.2.3. Discuss results of the weighted-based breast cancer prediction model

Several test scenarios were performed to verify the results shown in Figure 4.2. The scenarios involved removing one or more essential/non-essential risk factors to assess their impact on the accuracy of the optimizable tree-based classifier and the classification errors.

Figure 4.2 proves that the weighted version of the dataset outperforms the non-weighted one, with the performance increasing by 6.9% after weighting the risk factors. Similarly, Figure 4.3 shows that the importance of each risk factor varies in terms of its effect on defining the final risk degree.

The results indicate that the "Race" factor is the most influential, as the accuracy decreases by 4.3% after removing this factor. Other risk factors, such as age at first birth (agefirst), age group, Nrelbc, BMI, and Hispanic, also significantly affect performance when removed from the dataset.


FNR
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
AGE
RACE
NR
HT
SM
LM
BP
AF
BMI
HIS.
DENSITY
MENO.
AGE, RACE
RACE, NR
NR,AGE,RACE
MENO., BP, SM
Majority class FDR
Minor class FDR


FDR
45.00%
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
AGE
RACE
NR
HT
SM
LM
BP
AF
BMI
HIS.
DENSITY
MENO.
AGE, RACE
RACE, NR
NR,AGE,RACE
MENO., BP, SM
Majority class FDR
Minor class FDR

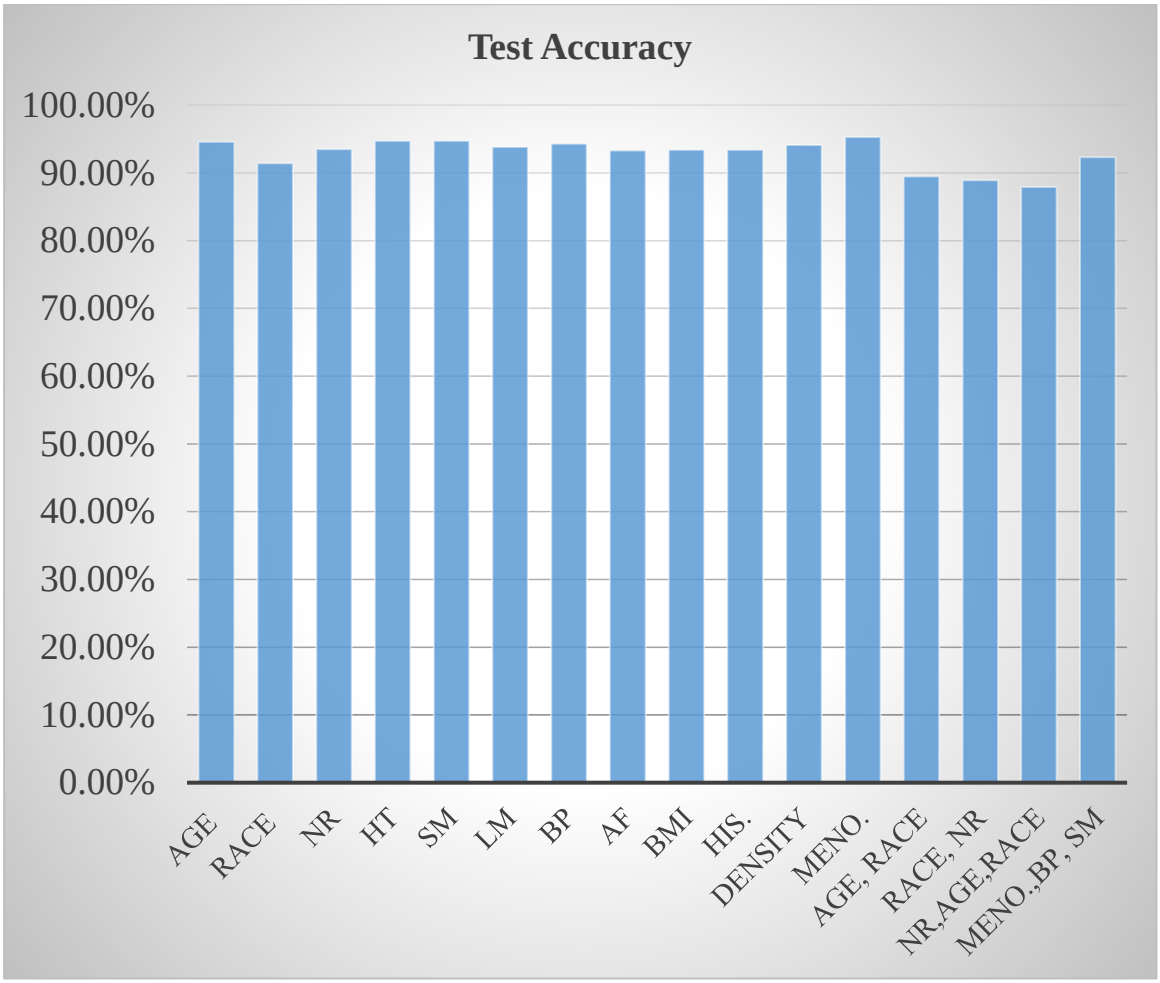
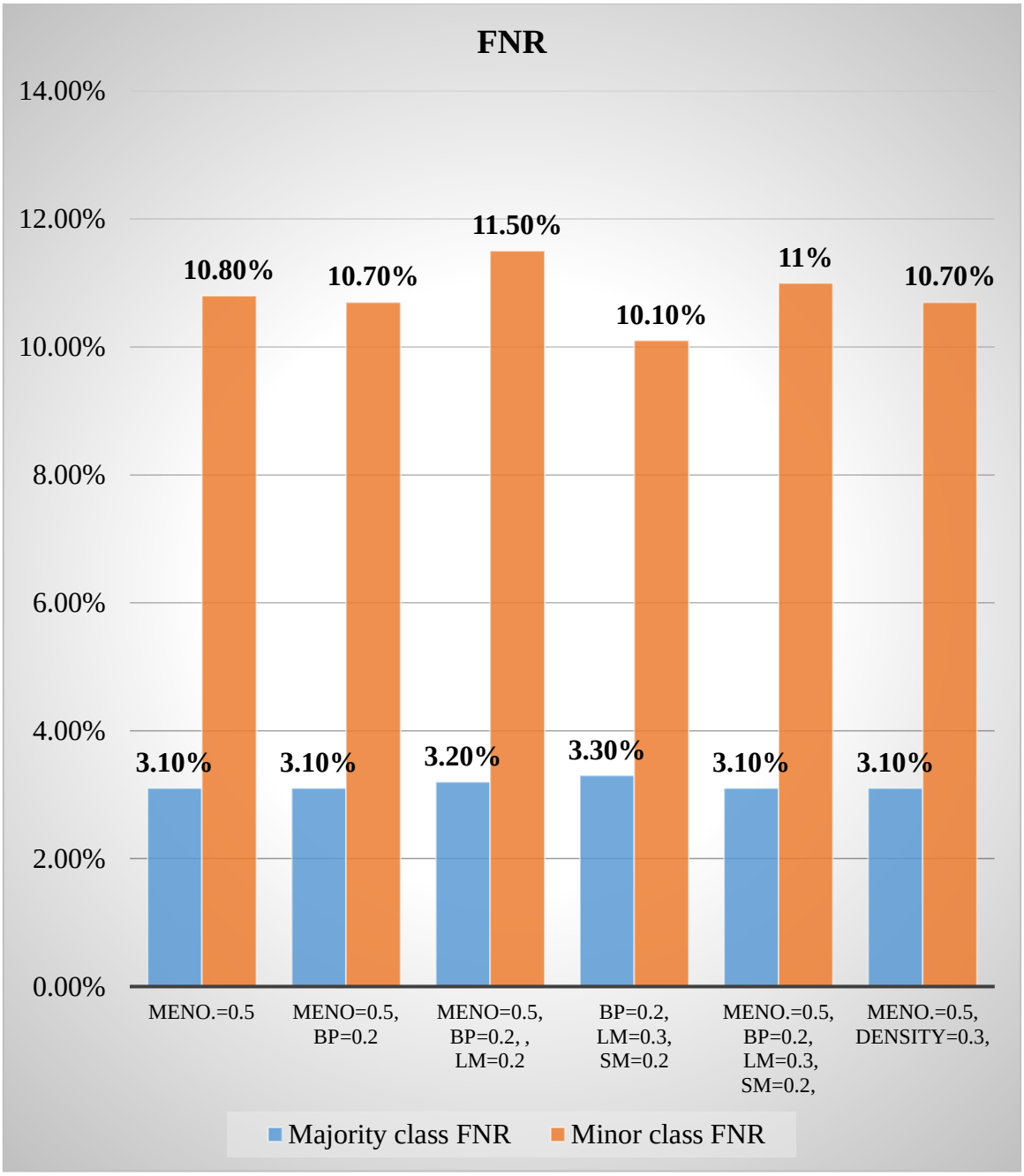


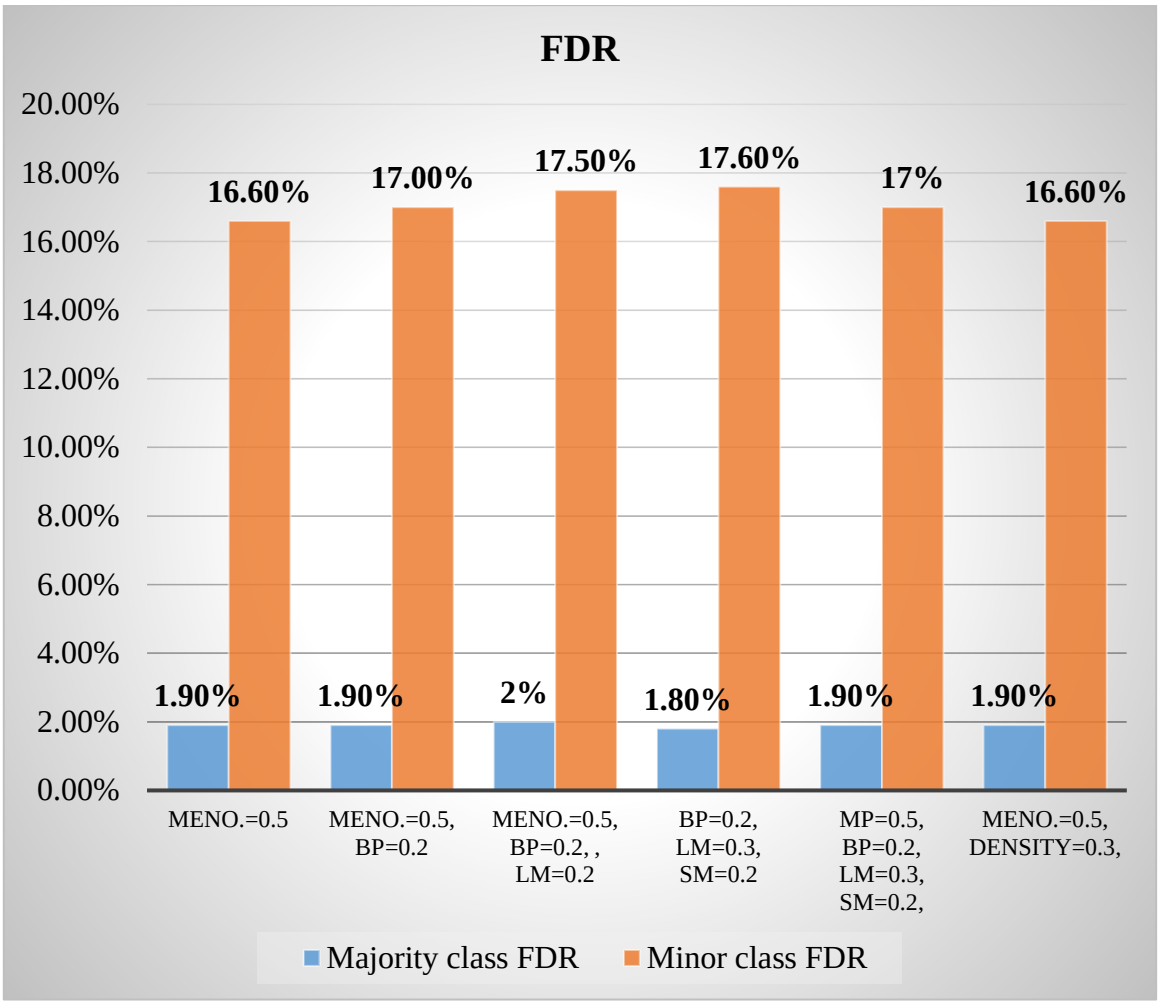
Figure 4.3. Evaluating the breast cancer prediction model under different risk factor combinations

Where: Menopause: meno., Hormone Therapy: HT, Surgmeno: SM, Lastmamo: LM, Brstproc: BP, Agefirst: AF, Hispanic: His., Nrelbc: NR.

Removing some risk factors, such as race, age group, agefirst, BMI, and Hispanic, results in an increase in the minor False Negative Rate (FNR). Furthermore, removing pairs of risk factors, such as (age and race) or (Nrelbc, age, and race), significantly degrades performance by 6.2% to 7.8%, and the minor class FNR error increases by 23% to 40%, indicating that these factors are essential. Conversely, risk factors such as menopause, surgical menopause (surgmeno), and hormone therapy only marginally decrease accuracy by 0.4% to 1%. Moreover, the absence of the three risk factors (menopause, brstproc, and surgmeno) results in a decrease in accuracy of only 3.4%.

Therefore, these factors have less impact than others on defining the final risk degree. To validate this conclusion, a down-weight approach was applied, where each weak-impact risk factor was assigned a weight of less than 1 (0.2, 0.3, 0.5, etc.), and the results are summarized in Figure 4.4.




FDR
20.00%
18.00%
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
16.60%
17.00%
17.50%
17.60%
17%
16.60%
1.90%
1.90%
2%
1.80%
1.90%
1.90%
MENO.=0.5
MENO.=0.5, BP=0.2
MENO.=0.5, BP=0.2, , LM=0.2
BP=0.2, LM=0.3, SM=0.2
MP=0.5, BP=0.2, LM=0.3, SM=0.2,
MENO.=0.5, DENSITY=0.3,
Majority class FDR
Minor class FDR

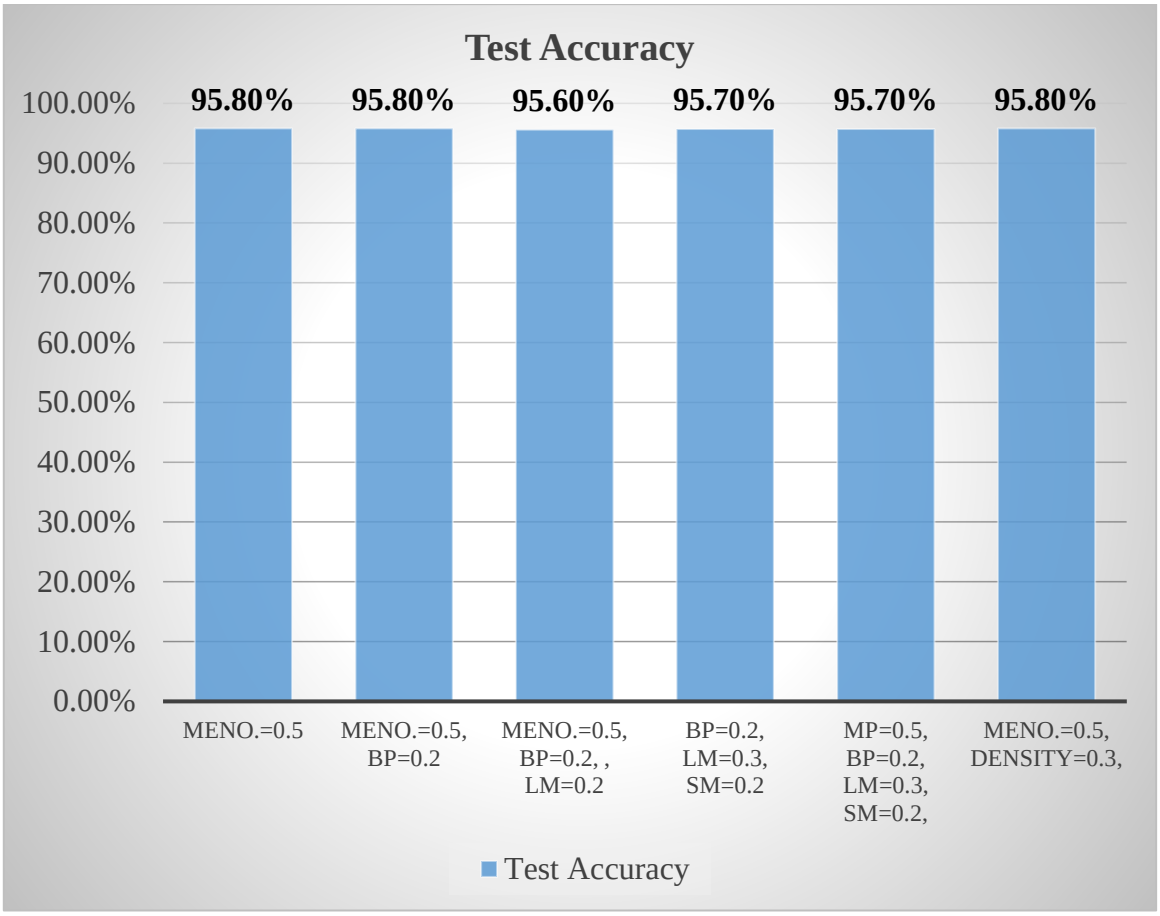


Figure 4.4. Effect of down-weighting the least essential risk factors on the performance of breast risk prediction model on the oversampled risk database

Scaling menopause, for example, by 0.5 ***improves*** the test accuracy by 0.1%, while scaling down other low-importance risk factors also ***improves*** the accuracy by 0.1% and ***reduces*** the False Negative Rate (FNR) error by 0.2%. However, in some cases, it increases the FNR of the minor class, mainly because the percentage of the minor class is small. Simultaneously, the False Discovery Rate (FDR) has decreased by 0.5-0.9%.

The scaling technique used on the oversampled dataset was also applied to the down-sampled and mixed datasets. Table 4.2 introduces a detailed comparison of how the choice of scaling factors (age=4, race=3, agefirst=2, nrelbc=3, current hormone therapy (current_hor)=3, menopause=0.5, density=0.3, brstproc=0.2, lastmamm=0.3, surgmeno=0.2) affects the performance of the three balanced datasets.

The results shown in Table 4.2 indicate that the down-sampled dataset has the highest accuracy (100%) and the lowest error rates (0%). However, it only contains 27.15% of the samples compared to the oversampled version.

Therefore, although the down-sampled dataset has the best accuracy, the oversampled and mixed versions perform better because they contain a larger number of samples, which allows for more accurate classification of new test samples.

Table 4.2. Performance of risk estimation model on three different balanced risk database

	Majority class FNR	Minor class FNR	Majority class FDR	Minor class FDR	Overall Validation Accuracy
Oversampling	**3.10%**	**10.70%**	**1.90%**	**16.60%**	**95.80%**
Down sampling	**0%**	0%	0%	0%	100%
Mixed	**3.20%**	8.90%	1.90%	14.40%	95.90%

4.3. Results of the range-based breast cancer prediction model

Two different training scenarios are performed in this section; the first scenario is done using a subset of the BCSC dataset, while the second scenario uses the enitre BCSC dataset. For both scenarios, the dataset is split into 80% training and 20% test.

4.3.1. Subset scenario

For each risk factor in the subset, the pre and post probabilities are computed using the 67633 records of the subset training dataset.

The post probabilities computed according to equation 5 are illustrated in Table 4.3.

Table 4.3. Cancer and non-cancer post probabilities of BCSC risk factors.

No.	Risk Factor	P(Prediction=Cancer |Innervalue ij)	P(Prediction=No cancer|Innervalue ij)
1	Menopause	Pre=78.34%, Post (age>55)= 30.29%, Unknown= 21.89%	Pre=21.66%, Post (age>55)= 69.71%, Unknown= 78.11%
2	Age group	35-39=4.67%; 40-44=11.81%; 45-49=22.47%; 50-54=41.57%; 55-59=29.2%; 60-64 =22.2%; 65-69=12.43%; 70-74=13.4%; 75-79=14.56%; 80-84=6.79%.	35-39=95.33%; 40-44=88.19%; 45-49=77.53%; 50-54=58.43%; 55-59=70.8%; 60-64 =77.8%; 65-69=87.57%; 70-74=86.6%; 75-79=85.44%; 80-84=93.21%.
3	Density	Almost entirely fatty: 9.99%, Scattered fibro-glandular: 45.88%, Heterogeneously dense: 52.68%, Extremely	Almost entirely fatty: 90.01%, Scattered fibro-glandular: 54.12%, Heterogeneously dense:

		dense: 38.24%, Unknown: 20.97%	47.32%, Extremely dense: 61.76%, Unknown: 97.03%
4	Race	White: 72.85% ; Asian/Pacific Islander: 36.36% ; Black: 10.62% ; Native American: 7.77% ; Other/mixed:28.1% ; Unknown: 19.66%.	White: 27.15% ; Asian/Pacific Islander: 63.64% ; Black: 89.38% ; Native American: 92.23% ; Other/mixed: 71.9% ; Unknown: 80.34%.
5	Hispanic	No: 28.33%;Yes: 81.6%; Unknown: 28.17%.	No: 71.67%;Yes: 18.4%; Unknown: 71.83%.
6	BMI	10-24: 18.94%; 25-29.99: 23.34%; 30-34.99: 31.41%; 35 or more: 41.58%; Unknown: 59.57%.	10-24: 81.06%; 25-29.99: 76.66%; 30-34.99: 68.59%; 35 or more: 58.42%; Unknown: 40.43%.
7	Age at first birth (agefirst)	Age<30: 30.54%; Age 30 or greater: 60.11%; Nulliparous: 60.24%; Unknown: 23.83%.	Age<30: 69.46%; Age 30 or greater: 39.89%; Nulliparous: 39.76%; Unknown: 76.17%.
8	Number of first degree relatives with breast cancer (nrelbc)	Zero: 21.75%; One: 49.02%; 2 or more: 96.99%; Unknown: 24.68%.	Zero: 78.25%; One: 50.98%; 2 or more: 3.01%; Unknown: 75.32%.
9	Previous breast procedure (brstproc)	No: 18.11% ; Yes:87.41%; Unknown: 36.34%.	No: 81.89% ; Yes:12.59%; Unknown: 63.66%.
10	last mammogram before the index mammogram (lastmamm)	Negative: 63.47%; False positive: 88.77%; Unknown: 20.25%.	Negative: 36.53%; False positive: 11.23%; Unknown: 79.75%.
11	Surgical menopause	Natural: 31.38%; Surgical: 81.27%; Unknown or not Menopausal: 32.57%.	Natural: 68.62%; Surgical: 18.73%; Unknown or not Menopausal: 67.43%.
12	Hormone therapy	No: 29.08%; Yes: 82.6%; Unknown: 31.92%.	No: 70.92%; Yes: 17.64%; Unknown: 68.08%.

The post probabilities of the column "count" is computed using the distribution of cancer and non-cancer classes as follows:

P(NCancer|count<2)=0.3, P(Cancer|count<2)=0.3.

P(NCancer|count>=2&count<50)=0.8, (Cancer|count>=2 & count<8)=0.8.

P(NCancer|count>=50&count<1000)=0.95, P(Cancer|count>=8 & count<16)=0.95.

P(NCancer|count>=1000)=1, P(Cancer|count>=16)=1.

In this step, we utilize the probabilistic statistics from the previous stage and the weights of the risk factors obtained from results of the first branch of the study (weighted-based model).

The objective of these calculations is to evaluate the probabilistic model and determine the final prediction scores, BCPScancer and BCPSN_cancer for all records of the dataset.

Figure 4.5 displays the distribution of the "result prediction score" of the subset dataset. Notably, the "non-cancer" class is divided into several subclasses ('19', '20', '21', '22', '23', '24', '25', '26', '27', '28'), which represent low-predicted percentages of breast cancer instead of using a single class to indicate the presence or absence of breast cancer.

In contrast, the "cancer class" is divided into 26 subclasses ('48', '49', '50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60', '61', '62', '63', '64', '65', '66', '67', '68', '69', '70', '71', '72', '73'), which represent high-predicted percentages of breast cancer scores.

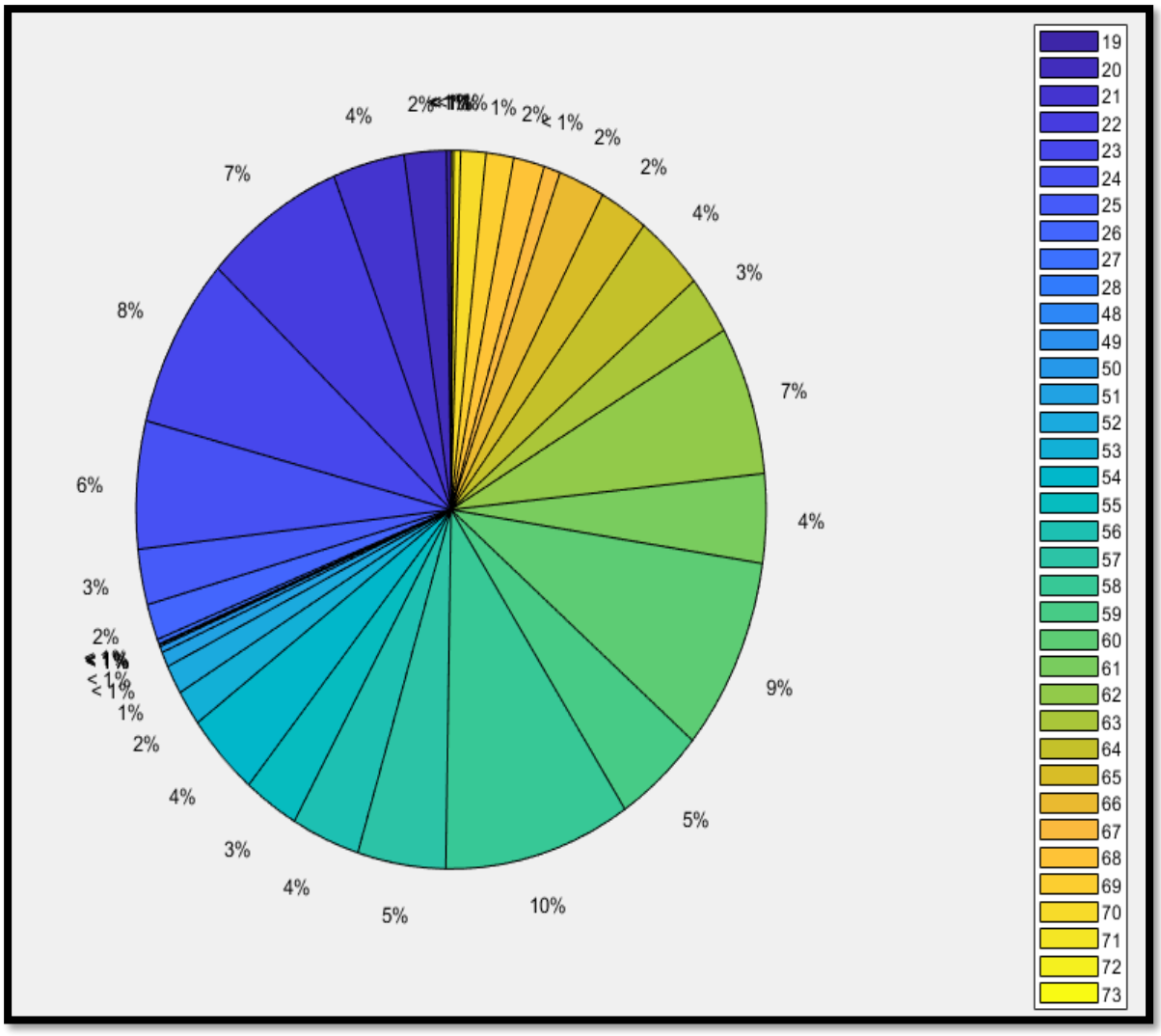


Figure 4.5. Distribution of the range-based breast cancer prediction categories in the subset-dataset scenario

4.3.2. Entire Dataset Scenario

The same experiments were conducted to determine the distribution of the subclasses of the entire BCSC dataset, as shown in Figure 4.6. The "cancer" class was divided into the same number of subclasses, but with a different distribution due to the different distribution of the "cancer" and "non-cancer" classes in the original dataset. Additionally, three subclasses ("29", "30", and "31") were identified for the "non-cancer" class. There is a notable difference in the distribution between the sub-dataset and the entire dataset. Figure 4.6 demonstrates this significant difference, where the "non-cancer" categories (ranging from "19" to "31") have higher percentages than the "cancer" categories. This difference is expected since the original dataset has almost 84% of its samples belonging to the "non-cancer" class.

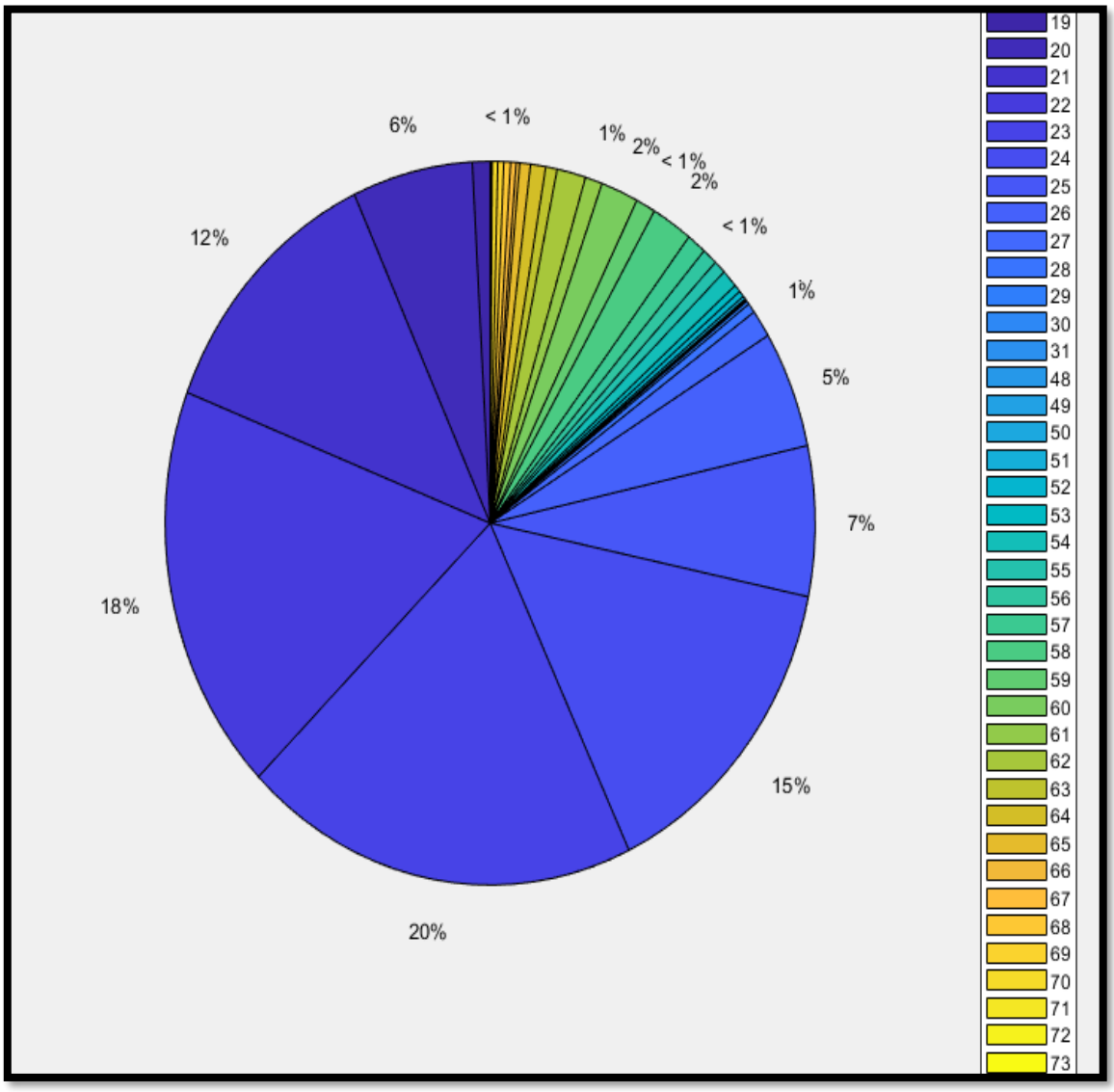
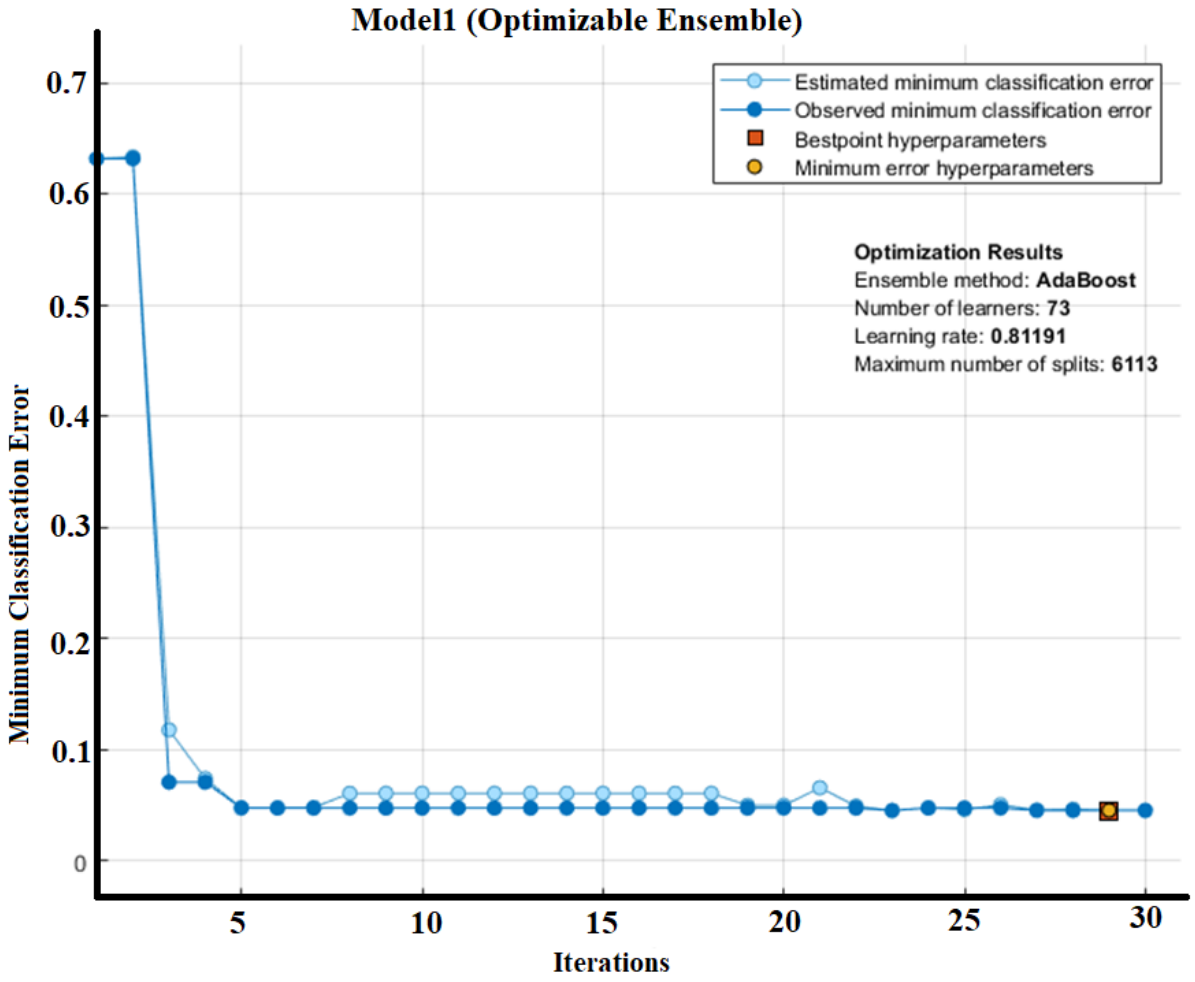


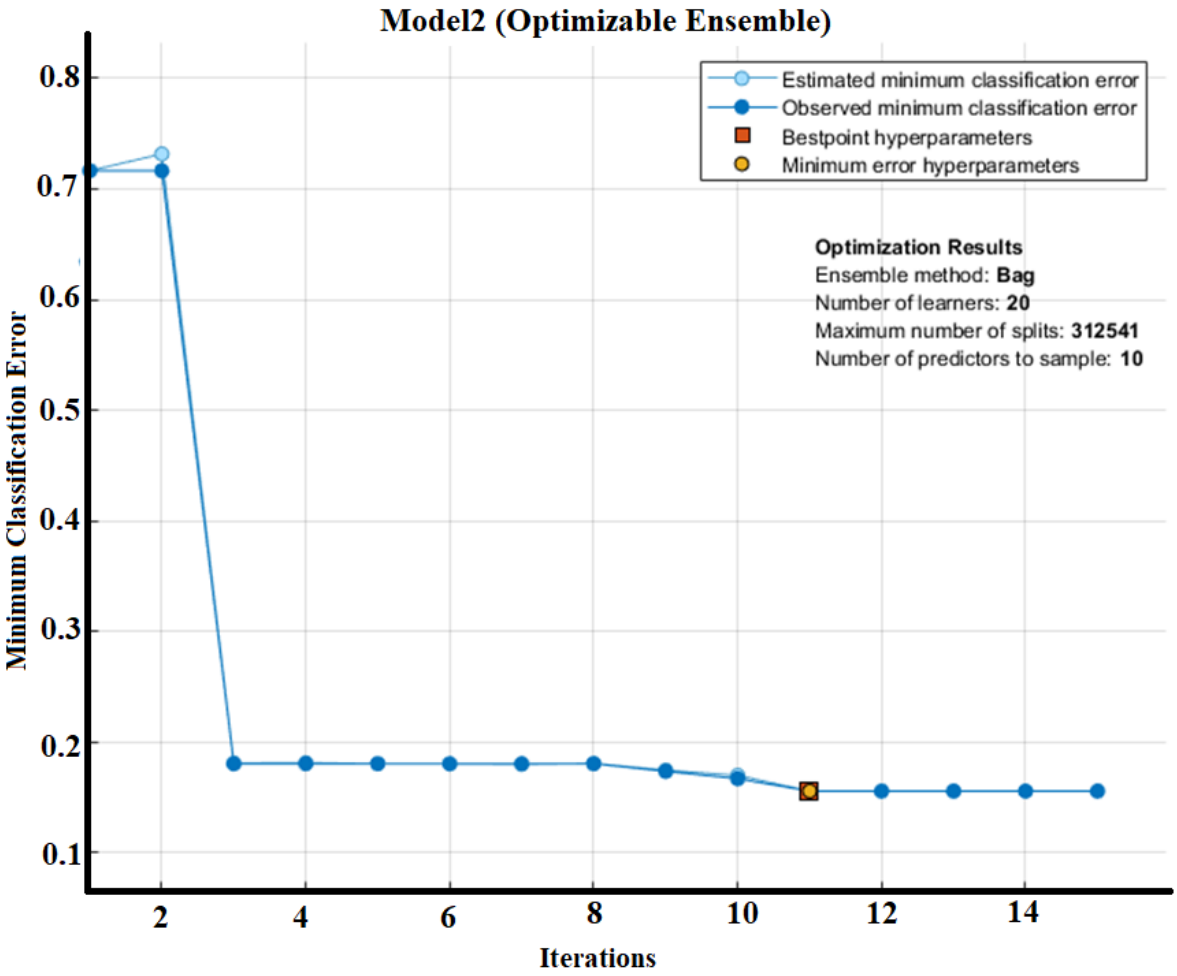
Figure 4.6. Distribution of the range-based breast cancer prediction categories in the entire-dataset scenario

4.3.3. Ensemble model training and evaluation

Two learned models were generated by feeding both the sub-dataset and the entire dataset into the ensemble classifier. Hyperparameters optimization was applied in both training scenarios using the AdaBoost ensemble method and Bayesian optimization. The Minimum Classification Error (MCE) curve of the training process was computed for both the sub-dataset and the entire dataset (Figure 4.7 shows those curves).



A) MCE of Subset dataset



B) MCE for Entire dataset

Figure 4.7. MCE of the trained range-based ensemble model

During the iterations from 10 to 30, the MCE value of the sub-dataset was consistently lower than that of the entire dataset by 0.1.

Many evaluation metrics are computed to evaluate the trained ensemble model that has been trained using the sub dataset and the entire dataset. Those metrics include TPR, FNR, PPR, FDR.

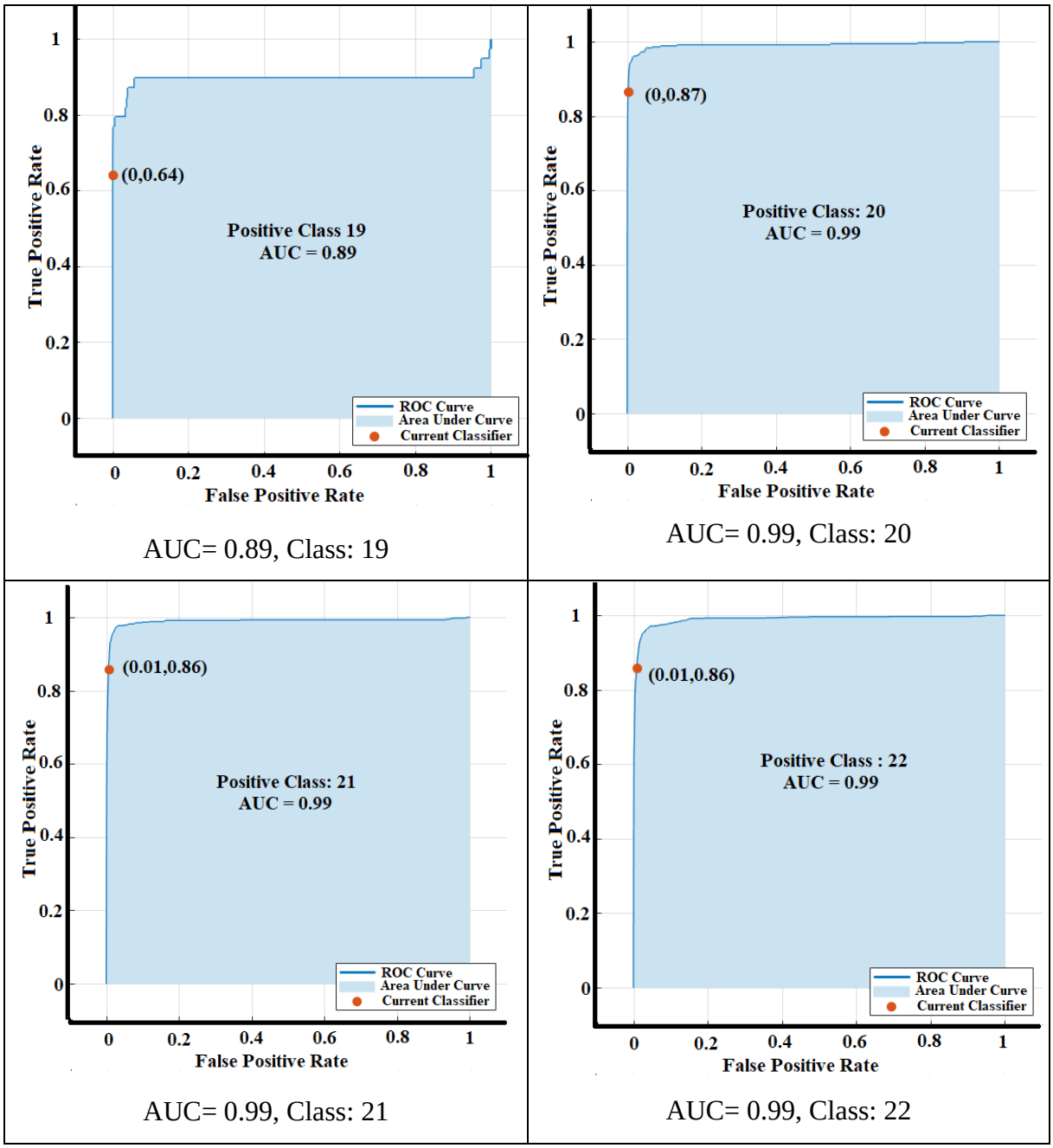
The evaluation results of the trained ensemble model using the range-based versions of the sub-dataset and the entire dataset are presented in Table 4.4.

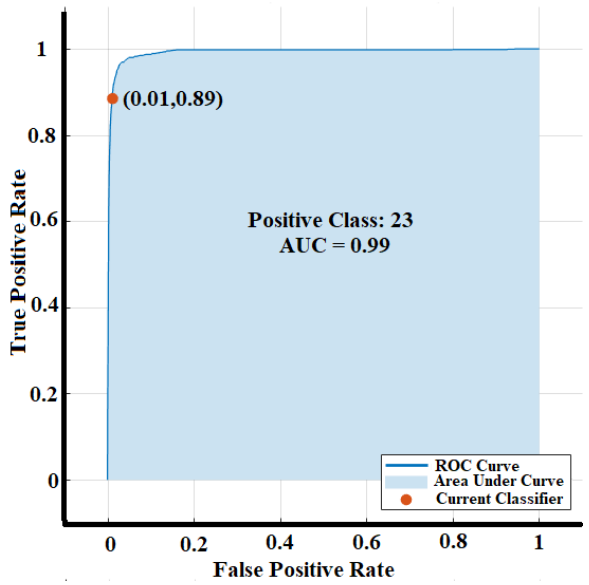
Table 4.4. Evaluation results of the ensemble model using the sub and whole dataset.

Class	TPR_sub %	TPR_Whl %	FNR_sub %	FNR_Whl %	PPR_sub %	PPR_Whl %	FDR_sub %	FDR_Whl %
19	64.1	77.9	35.9	22.1	83.33	86.7	16.7	13.3
20	86.58	83.9	13.42	16.1	87.53	87.5	12.47	12.5
21	85.85	84.4	14.15	16.6	82.41	85.3	17.59	14.7
22	85.88	84.4	14.12	16.6	88.12	85.6	11.88	14.4
23	88.61	83.9	11.39	16.1	87.8	85.9	12.2	14.1
24	87.14	82.9	12.86	17.1	87.41	85.2	12.59	14.8
25	77.73	75.3	22.27	24.7	77.73	80.6	22.27	19.4
26	84.62	83.9	15.38	16.1	83.4	85.3	16.6	14.7
27	77.78	72.3	22.22	27.7	87.5	78.6	12.5	21.4
28	75	64.3	25	35.7	75	72.3	25	27.7
29	-	64.5	-	35.5	-	69.6	-	30.4
30	-	69.4	-	30.6	-	73.5	-	26.5
31	-	0	-	100	-	-	-	100
48	100	100	0	0	100	100	0	0
49	100	100	0	0	100	94.1	0	5.9
50	100	100	0	0	100	82.1	0	17.9
51	100	100	0	0	95.9	86	4.1	14.0
52	100	100	0	0	100	81.9	0	18.1
53	98.16	100	1.84	0	100	83.6	0	16.4
54	100	96.6	0	3.4	100	83.1	0	16.9
55	100	99.2	0	0.8	100	84.0	0	16.0
56	100	100	0	0	99.16	85.8	0.84	14.2
57	100	99.6	0	1.4	100	86.8	0	13.2
58	99.38	99.2	0.62	0.8	100	86.4	0	13.6
59	100	100	0	0	98.73	84.9	1.27	15.1
60	99.66	98.6	0.34	1.4	100	84.8	0	15.2
61	100	99.3	0	0.7	100	87.2	0	12.8
62	100	100	0	0	100	87.1	0	12.9
63	100	99.1	0	0.9	100	85.3	0	14.7
64	100	99.3	0	0.7	100	85.8	0	14.2
65	100	98.1	0	1.9	100	89.8	0	10.2
66	100	100	0	0	100	89.5	0	10.5
67	100	100	0	0	100	83.4	0	16.6
68	100	100	0	0	98.17	93.0	1.83	7.0
69	97.92	100	2.08	0	97.9	90.1	2.1	9.9
70	97.7	100	2.3	0	100	90.4	0	9.6
71	100	100	0	0	100	90.5	0	9.5
72	100	100	0	0	100	79.2	0	20.8
73	100	100	0	0	100	100	0	0

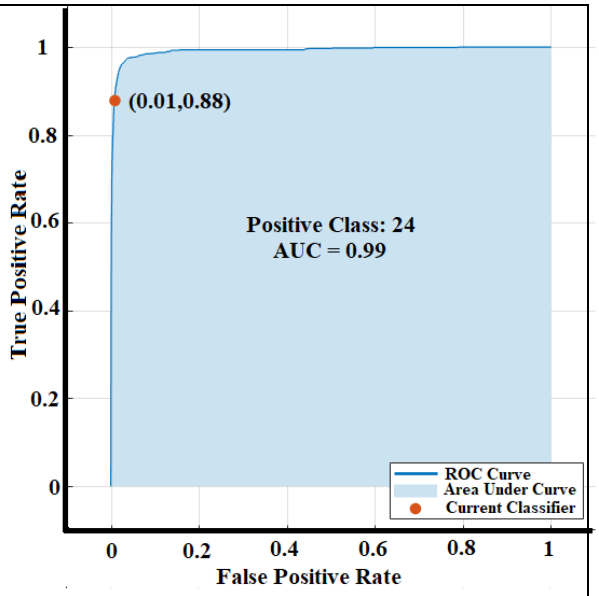
According to the statistics presented in Table 4.4, the average TPR for the sub-dataset and entire dataset are 94.61% and 92.52%, respectively. Similarly, the average PPR for the sub-dataset and entire dataset are 92.28% and 85.55%, respectively. The total accuracy for the sub-dataset and entire dataset is 95.5% and 85.3%, respectively.

To assess the ability to distinguish between different subclasses, the AUC is utilized for all trained ensemble models. Detailed AUC results for all subclasses ("19"-"73") of the sub-dataset and entire dataset are presented in Figure 4.8.

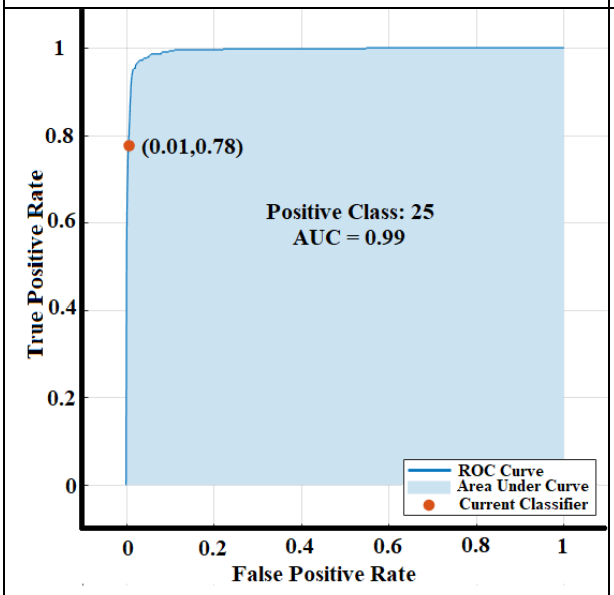




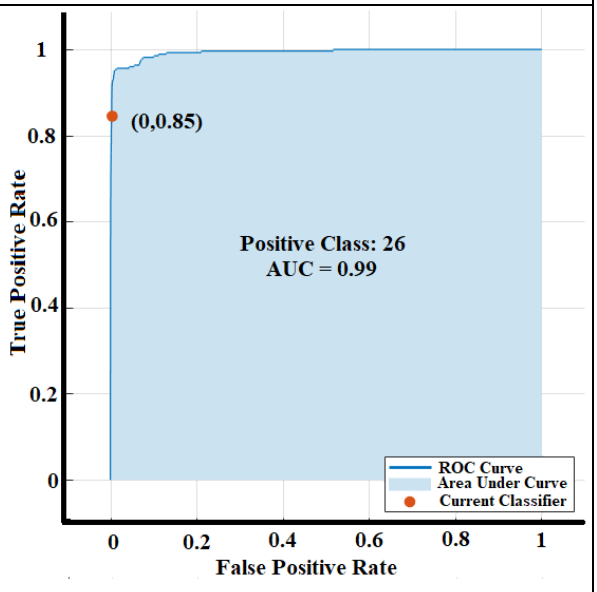
AUC= 0.99, Class: 23



AUC= 0.99, Class: 24



AUC= 0.99, Class: 25



AUC= 0.99, Class: 26

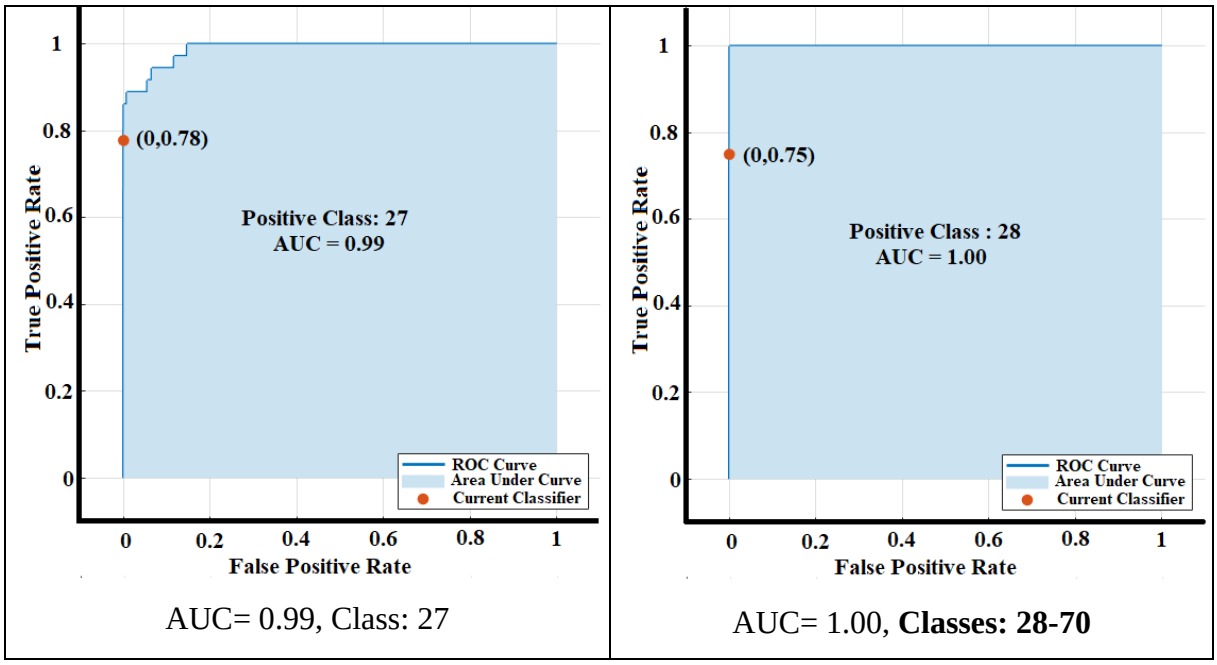


Figure 4.8. AUC and ROC curves of the entire categories of the BCSC dataset

For both the sub-dataset and entire dataset, all "cancer" subclasses have an AUC of 1. However, some "non-cancer" subclass has the low AUC value (like class 19). This finding is supported by Table 4.4, which shows that subclass "19" has low accuracy and high error rates.

There are other classes like class "31" which also has a low accuracy. The reason for this issue is that subclasses like "19" or "31" have a significantly smaller number of samples (for example class "31" has only 7 samples for training and 3 for validation) compared to the other subclasses. This limited sample size may have resulted in less accurate training of the model for this particular subclass, leading to lower AUC and higher FNR and FDR values.

4.3.4. Variance discussion

The new subclasses introduced in the modified BCSC dataset are associated with their original containing class, meaning that subclasses "19" to "31" belong to the "non-cancer" class, while subclasses "48" to "73" belong to the "cancer" class. To accurately represent the results of the modified model, we conducted performance evaluations with two additional trials: one with a ±1 class-variance tolerance and another with a ±2 class-variance tolerance.

The closely related subclasses (±1 or ±2) have similar cancer/non-cancer scores and can be treated as a single subclass. For example, if the actual subclass is "21," the accepted true classes for ±1 class-variance are "20," "21," and "22," while for ±2 class-variance, the accepted classes are "19," "20," "21," "22," and "23." In the first trial, two biases of the main classes are

allowed, so if the sample has the original true class "i," the expected valid classes are (i-1, i, i+1). In the second trial, the expected valid classes are (i-2, i-1, i, i+1, i+2).

Tables 4.5 and 4.6 show the detailed results of these two trials for both the sub-dataset and entire dataset, respectively.

The results indicate that the average TPR of the original confusion matrix of the sub-dataset is 90.1564%, while it increases by 4.2% and 5.38% for the ±1 and ±2 variance scenarios, respectively. Similarly, the PPR of the ±1 and ±2 variance scenarios have been enhanced by 4.56% and 4.72%, respectively.

Likewise, the average TPR of the original confusion matrix of the entire dataset is increased by 8.66% and 8.76% for both the ±1 and ±2 variance scenarios, respectively (refer to Table 4.6). The average PPR values of the ±1 and ±2 variance scenarios also increased by 5.33% and 5.55%, respectively.

The accuracy computation also supports the same conclusion, where the original accuracy was 85.3%, but it increases by 5.82% and 6.03% for the ±1 and ±2 class-variances, respectively.

Table 4.5. Variance results (TPR and PPR) of the sub dataset

Class No.	Original TPR	TPR (±1)	TPR (±2)	Original PPR	PPR (±1)	PPR (±2)
19	64.1	100	100	83.33	100	100
20	86.58	100	99.72	87.53	99.72	100
21	85.85	100	99.85	82.41	99.85	100
22	85.88	99.8	100	88.12	100	100
23	88.61	99.9	100	87.8	100	100
24	87.14	100	99.69	87.41	99.69	100
25	77.73	100	100	77.73	100	100
26	84.62	99.6	100	83.4	100	100
27	77.78	100	100	87.5	100	100
28	75	100	100	75	100	100
48	100	100	100	100	100	100
49	100	100	100	100	100	100
50	100	100	100	100	100	100
51	100	100	95.9	95.9	95.9	100
52	100	100	100	100	100	100
53	98.16	100	100	100	100	100
54	100	100	100	100	100	100
55	100	100	100	100	100	100

56	100	100	99.16	99.16	99.16	100
57	100	100	100	100	100	100
58	99.38	99.7	100	100	100	100
59	100	100	100	98.73	100	100
60	99.66	100	100	100	100	100
61	100	100	100	100	100	100
62	100	100	100	100	100	100
63	100	100	100	100	100	100
64	100	100	100	100	100	100
65	100	100	100	100	100	100
66	100	100	100	100	100	100
67	100	100	100	100	100	100
68	100	100	100	98.17	100	100
69	97.92	100	100	97.9	100	100
70	97.7	100	100	100	100	100
71	100	100	100	100	100	100
72	100	100	100	100	100	100
73	100	100	100	100	100	100

Table 4.6. Variance results (TPR and PPR) of the entire dataset

Class No.	Original TPR	TPR (±1)	TPR (±2)	Original PPR	PPR (±1)	PPR (±2)
19	77.9	98.28	98.28	86.7	86.7	86.7
20	83.9	97.3	97.34	87.5	99.98	100
21	84.4	97.54	97.54	85.3	99.89	100
22	84.4	97.49	97.51	85.6	99.7	99.86
23	83.9	97.02	97.11	85.9	99.91	99.95
24	82.9	96.96	97.17	85.2	99.64	99.81
25	75.3	96.99	97.09	80.6	99.8	99.88
26	83.9	97.39	97.28	85.3	99.39	99.71
27	72.3	96.32	96.63	78.6	100	100
28	64.3	92.46	95.65	72.3	99.35	100
48	64.5	96.77	96.77	69.6	99.13	100
49	69.4	97.22	97.22	73.5	94.12	100
50	100	100	100	82.1	82.1	82.1
51	100	100	100	86	86	86
52	100	100	100	81.9	81.9	81.9
53	100	100	100	83.6	83.6	83.6
54	96.6	100	100	83.1	83.1	83.1

55	99.2	99.32	99.32	84	85.16	85.16
56	100	99.6	99.6	85.8	85.8	85.8
57	99.6	99.6	99.6	86.8	86.8	86.8
58	99.2	99.2	99.2	86.4	86.4	86.4
59	100	100	100	84.9	84.9	84.9
60	98.6	98.6	98.6	84.8	84.8	84.8
61	99.3	99.3	99.3	87.2	87.2	87.2
62	100	100	100	87.1	87.1	87.1
63	99.1	99.1	99.1	85.3	85.3	85.3
64	99.3	99.3	99.3	85.8	85.8	85.8
65	98.1	98.1	98.1	89.8	89.8	89.8
66	100	100	100	89.5	89.5	89.5
67	100	100	100	83.4	83.4	83.4
68	100	100	100	93	93	93
69	100	100	100	90.1	90.1	90.1
70	100	100	100	90.4	90.4	90.4
71	100	100	100	90.5	90.5	90.5
72	100	100	100	79.2	79.2	79.2
73	100	100	100	100	100	100

4.4. Results of the third branch (ML and DL fusion model)

The proposed method is evaluated through several training and test scenarios in the experimental section. For the Bi-LSTM model, five distinct test scenarios are implemented. Three of these scenarios involve modifying the Bi-LSTM architecture by changing the number of neurons and learning epochs, while the remaining two scenarios relate to the training and test percentages.

To evaluate these five scenarios, performance evaluation metrics such as True Positive Rate (TPR), Positive Predictive Rate (PPR), False Negative Rate (FNR), False Discovery Rate (FDR), and test accuracy are utilized. Table 4.7 includes the evaluation results of the Bi-LSTM model.

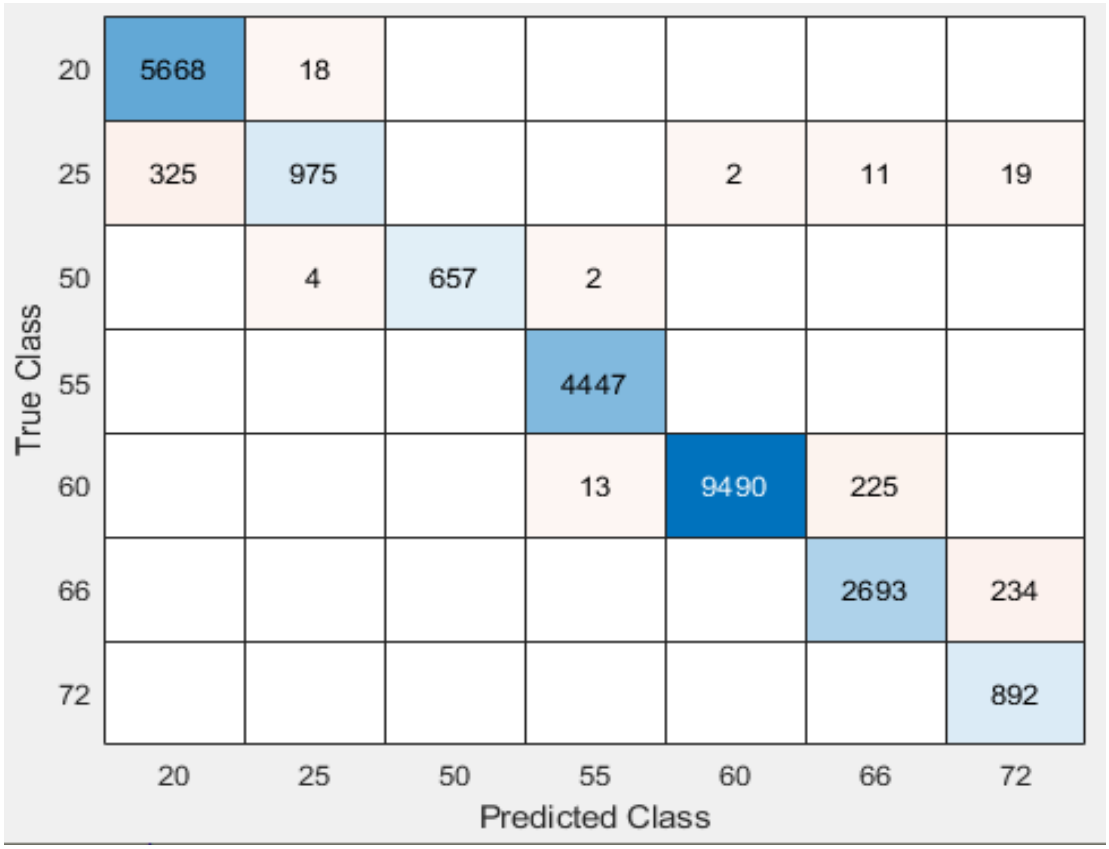
Table 4.7. Results of different test scenarios of Bi-LSTM model

Scenario	TPR	FNR	PPR	FDR	Test Accuracy
Bi-LSTM (100 iterations, 300 neurons)	65.1148%	34.88%	91.49%	8.509%	86.84%
Bi-LSTM (150 iterations, 300 neurons)	94.51%	5.49%	95.55%	5.54%	96.68%
Bi-LSTM (150 iterations, 400 neurons)	88.979%	11.02%	96%	4%	95.197%
Bi-LSTM (150 iterations, 300 neurons, test Percentage= 30%)	92.05%	7.95%	93.11%	6.89%	91.38%
Bi-LSTM (150 iterations, 300 neurons, test Percentage= 35%)	91.7116%	8.28%	96.814%	3.186%	93.85%

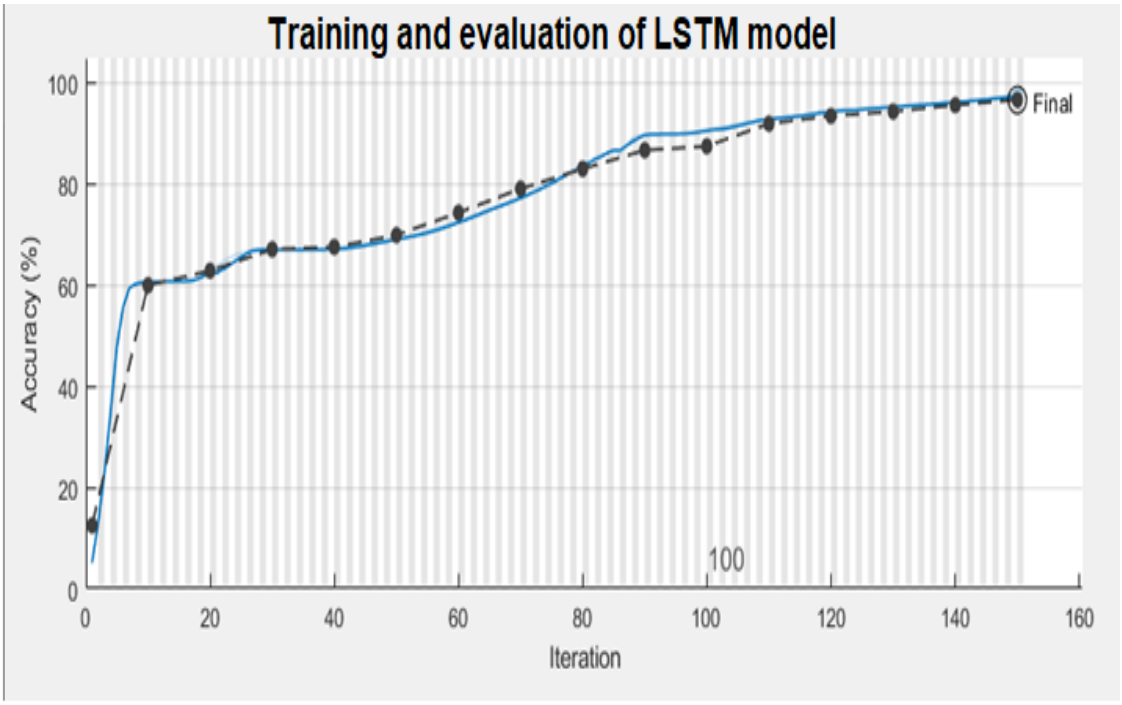
Table 4.7 demonstrates that the optimal Bi-LSTM architecture is achieved by using 300 neurons and 100-150 iterations for training. Regarding the data splitting, the best-case scenario is obtained by using 20% of the dataset samples as a test set. The last two scenarios reveal that by increasing the number of samples in the test set, the performance decreases.

Figure 4.9 shows the confusion matrix of the best DL model. The figure illustrates that the highest FNR error rate is associated with class "25" with FNR = 0.26, while the best TPR is linked to class "55" with TPR = 100%. Furthermore, class "50" has the highest PPR value at 100%, while class "72" has the worst FDR value of 0.779.

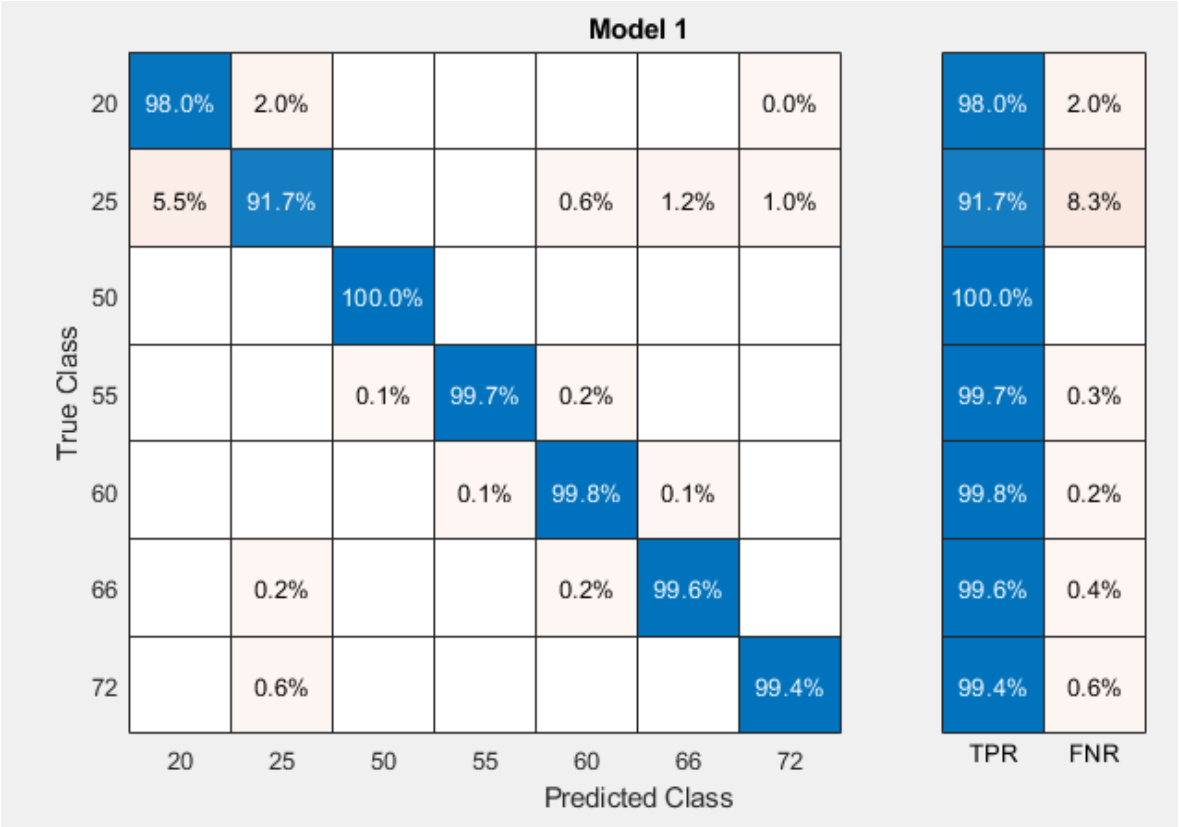
For the ensemble learning ML model, the minimum classification error of the boosted decision tree models is 1.1%, and the confusion matrix with TPR, FNR, PPR, and FDR is shown in Figure 4.9. The figure reveals that the best TPR is related to class "50," and the best PPR corresponds to class "55." These results are consistent with those obtained by the DL model.



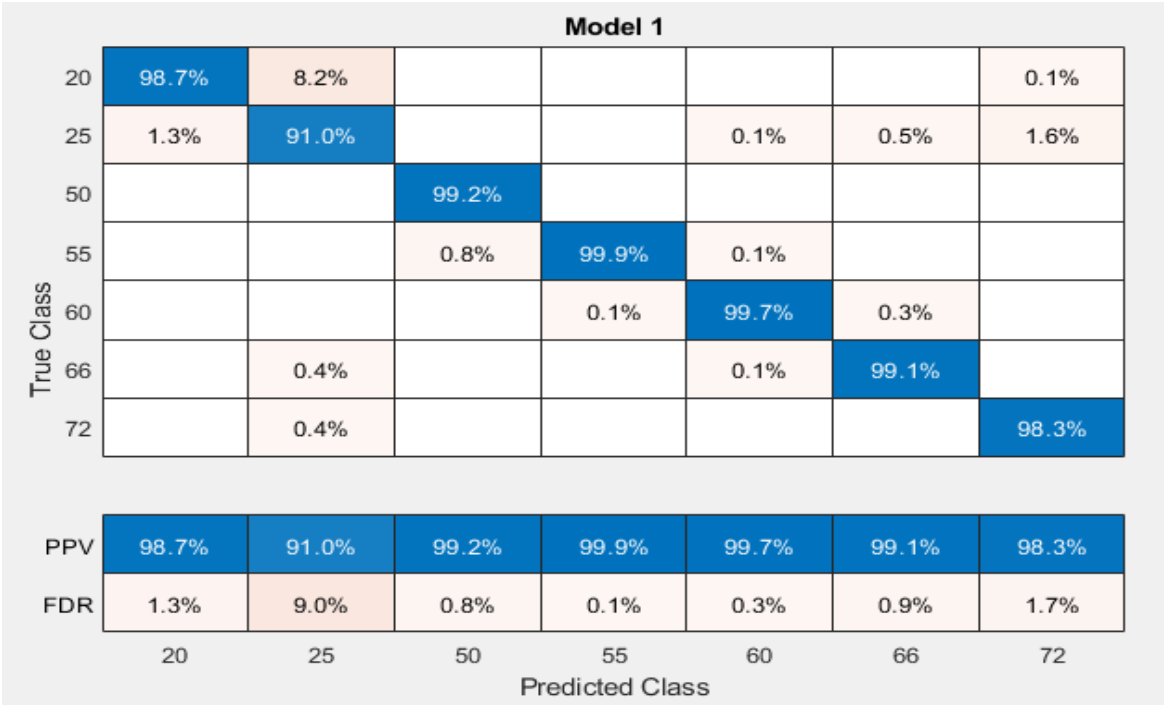
A) Confusion Matrix



B) Training and validation progress



C) TPR



D) PPR

Figure 4.9. Confusion matrix and performance metrics (accuracy and loss) of the best ML ensemble scenario

4.4.1. Fusion of DL and ML models

The final test scenario involves fusing the scores of the DL and ML models, and a detailed comparison between the individual models and the fusion model is presented in Figure 4.10.

As demonstrated in Figure 4.10, the fused model outperforms the individual models. The accuracy is increased by 1.08% and 3.3% compared to the ML and DL individual models, respectively. Additionally, the TPR is improved by 1.66% and 5.46%, while the PPR is increased by 2.01% and 5.44% compared to the DL and ML individual models.

These results confirm that fusing the DL and ML models significantly enhances performance.

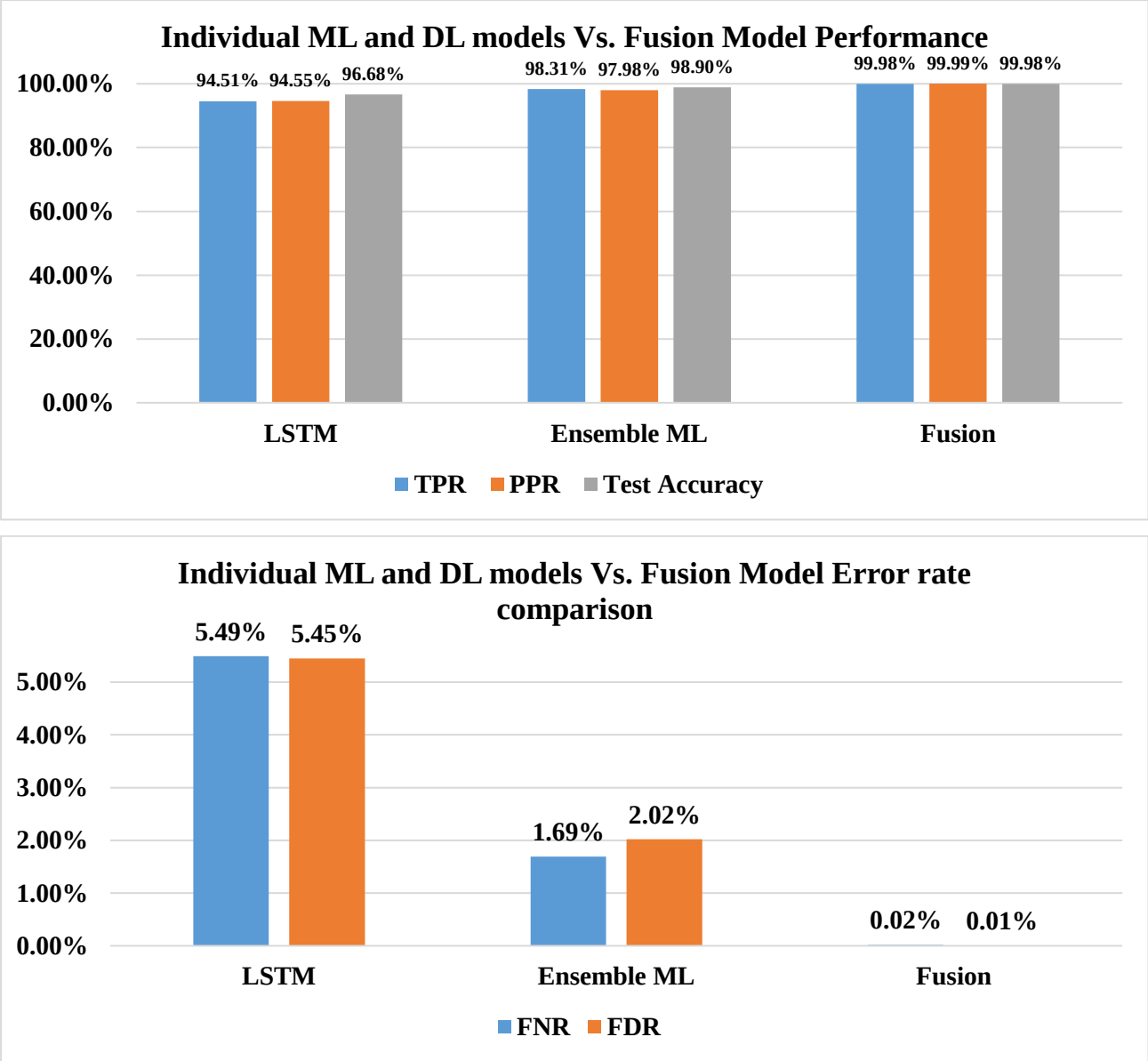


Figure 4.10. Performance comparison between individual and fusion model

Figure 4.10 proves that the fusion model is the best one since the error rates are almost 0% and the accuracy is 99.98%.

4.5. Results of the fourth branch of the study

In this part, the probabilistic model that have been built is reused and applied to the original entire dataset (without any balancing) to get all risk values as a range-based score. Now the target column (risk factor) is a value between 19% (the lowest value produced by the probabilistic model) and 70% (the highest value produced by the probabilistic model).

As mentioned in the methodology part, this branch will be performed using the specific values of the target column (risk scores) as the main classes of our problem, meaning that the problem will be a classification task.

However, the main issue in the classification thread is that there are many classes (39 classes) with different number of samples, and the balancing has little effect since there are classes with more than 50000 samples while others contain less than 50 samples. For this reason, the adjacent categories (classes) are merged to constitute unified categories and minimize the difference between classes since the adjacent categories represent a closed cancer prediction.

4.5.1. Results of the fourth branch (the classification thread)

After merging categories together, we got the following new classes of our classification problem illustrated in Table 4.8.

Table 4.8. Breast cancer new classes after merging the adjacent categories

Category (class)	Number of samples
20	60638
23	166873
27	41813
30	2031
50	1432
60	5629
70	2244

Then, the dataset is split into train and test (80% train, 20% test). The training set is balanced using SMOTE algorithm (oversampling) using a specific condition by which all classes with less than 2000 samples increased to 5000, and each class with number of samples less than 1000 is also increased to 10000. The oversampling uses the nearest two neighbors in generating the new samples.

After that, three ML models (RF, DT, LGBM, and Ensemble of the three models) and two DL models (1D-CNN and LSTM) are trained using the training set of the result dataset and then evaluated using the test set.

Table 4-9 includes the evaluation results of the trained ML and DL models.

Table 4.9 Evaluation results of the trained ML and DL models of the fourth scenario

Model	Accuracy %	Precision %	Recall %	F1-score %
LGBM	94.19	94	94	94
DT	90.83	91.13	90.83	90.97
RF	93.35	93.09	93.35	92.97
Ensemble (LGBM, DT, RF)	94.35	94.1	94.35	94.03
1D-CNN	93.12	90.46	93.12	91.7
LSTM	93.63	93.22	93.63	93.19
Ensemble of 1D-CNN and LSTM	94.35	94.1	94.35	93.62

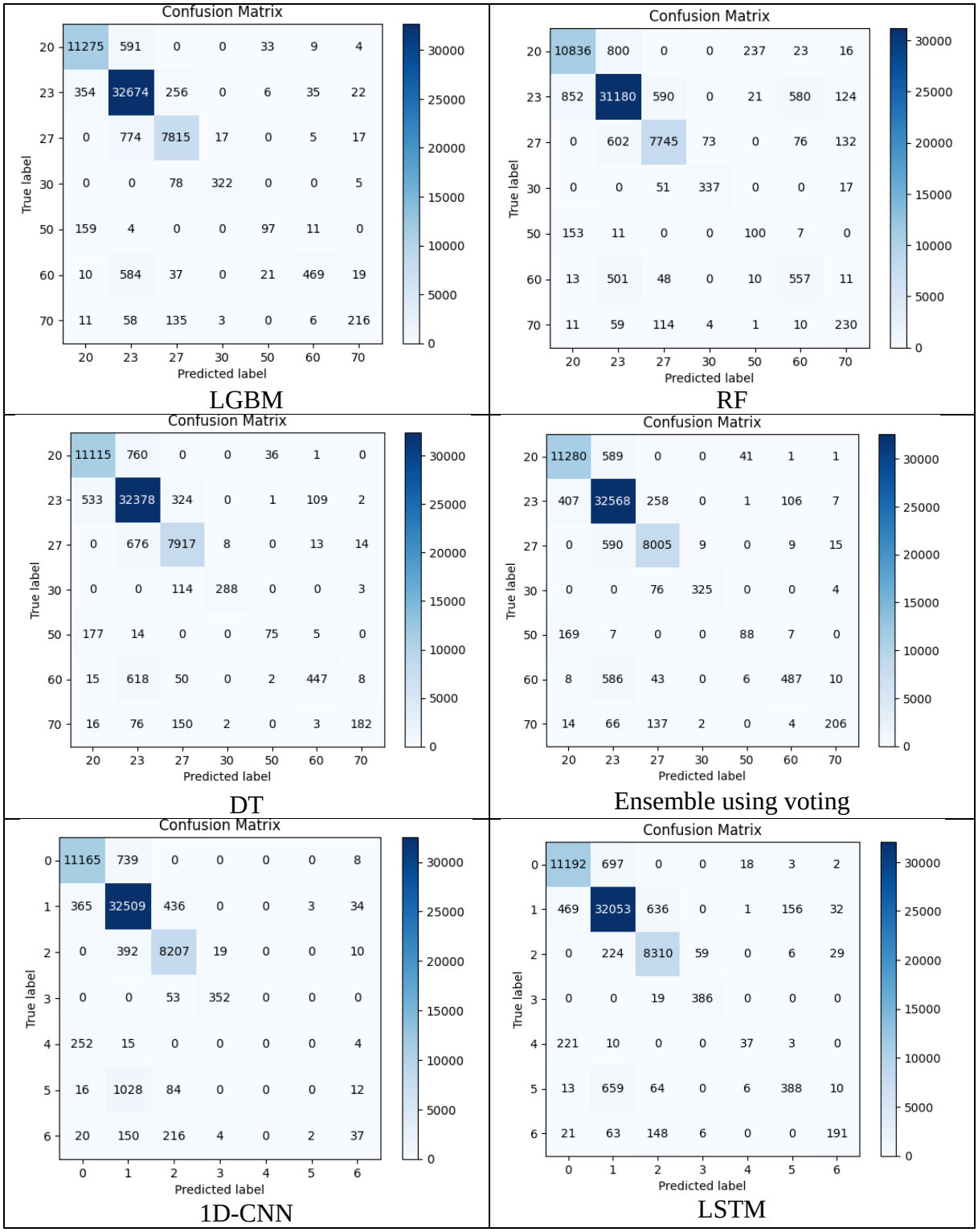
Table 4.9 shows that the ensemble ML and DL models achieved the best performance in terms of accuracy, precision, recall and F1-score. The best obtained accuracy is 94.35%.

To know the main reason of errors in the evaluation results, the confusion matrixes are also derived and shown in Figure 4.13.

As illustrated in Figure 4.11, the confusion matrixes of all models achieve best results for the classes with large number of samples (20, 23, and 27). However, the rest of classes which have little number of samples suffers from errors (false positive errors and false negative errors). The Ensemble ML and DL models' confusion matrixes have less error rates than the individual models.

Class "50" is the class with the highest error rates although the number of errors of this class are less than others, and this is due to the low number of samples of this class. Although class

"20" has 632 false negative errors (in case of ML ensemble model), but its precision and recall are 94.97% and 94.83%, respectively, and this is due to the large number of samples (there are 11280 true positive samples of this class).



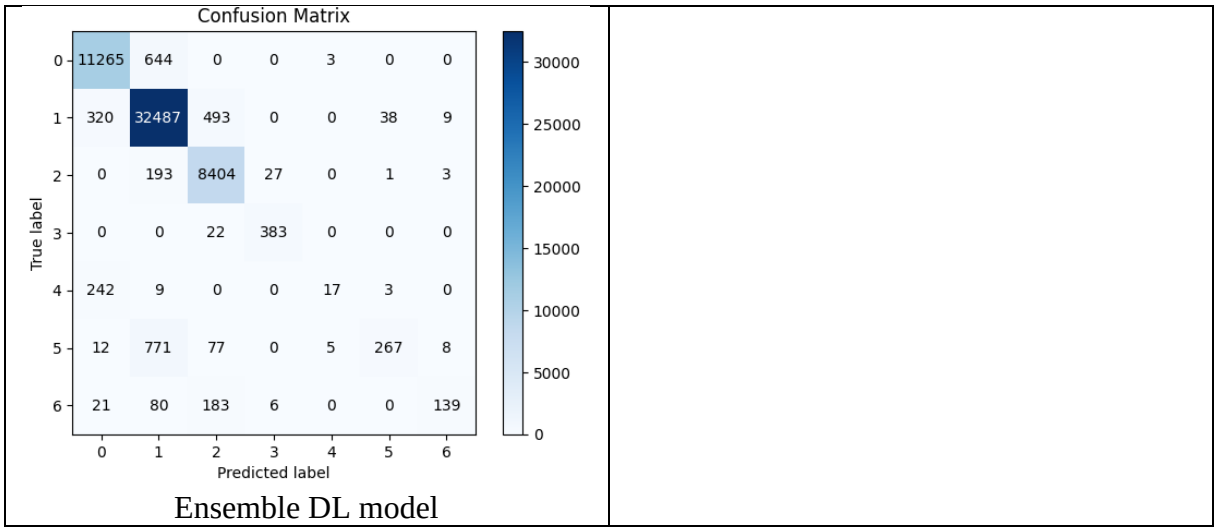


Figure 4.11. Performance evaluation of the trained ML and DL models of the fourth scenario (classification thread)

The difference between distribution of breast cancer score of the prediction results is shown in Figure 4.12.

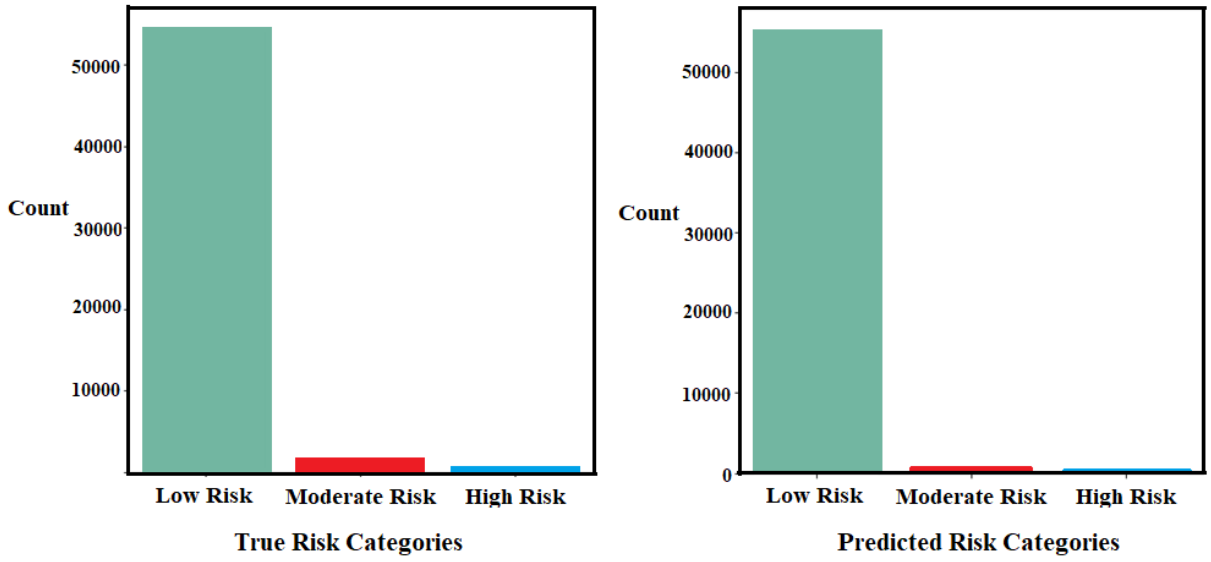
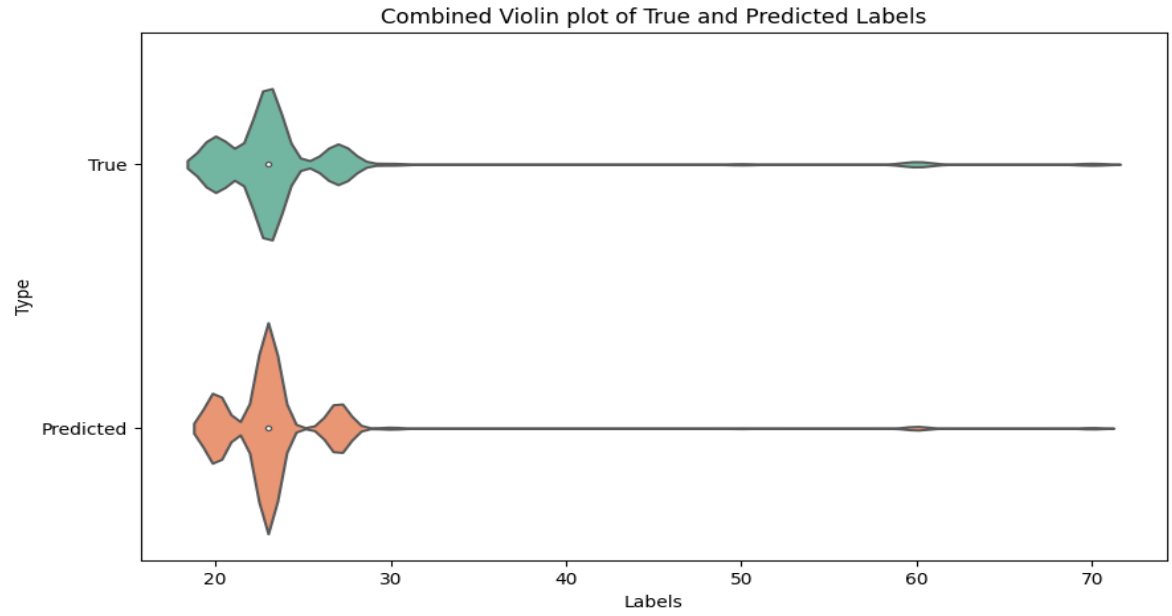


Figure 4.12. Distribution of the predicted breast cancer score

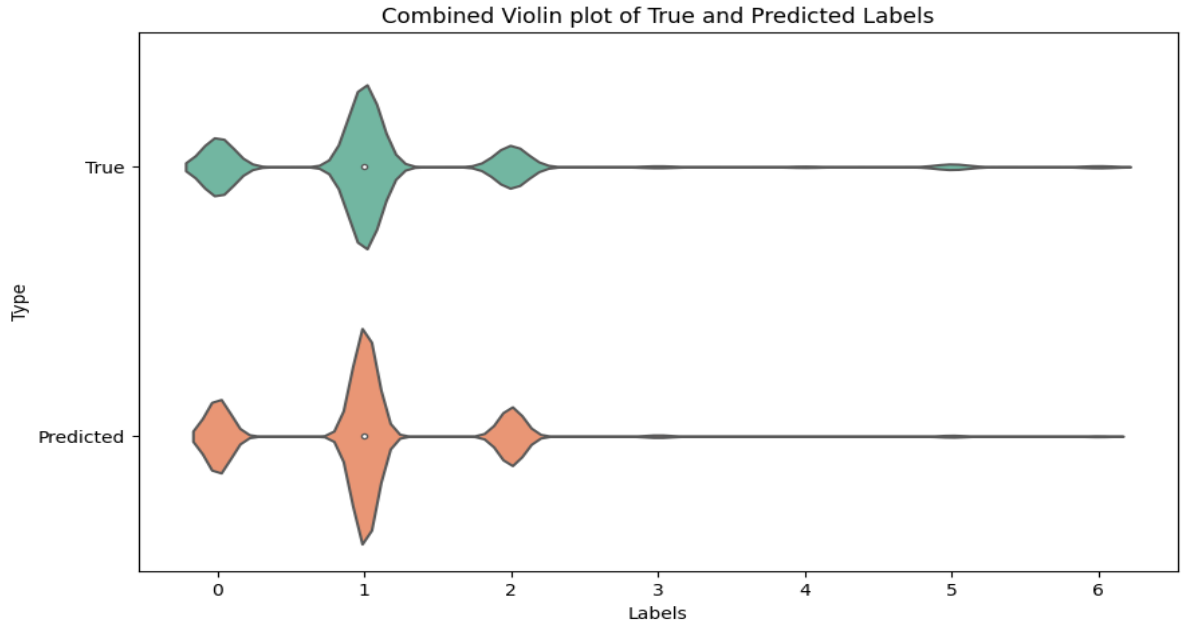
The distribution in Figure 4.12 proves that our task of predicting the right class among different size of them is a very hard one (although the balancing of the training set since the test set is preserved without any balancing so it will be biased to the non-cancer categories which are the categories with the greatest number of samples).

4.5.1.1. Violin Distribution Analysis

Now in terms of the distribution of the original and predicted cancer risk samples, we derived the plot in Figure 4.13-A. The distribution of the predicted and the original range-based cancer score proves the high accuracy of the proposed models since the two distributions are too closed. The same conclusion is confirmed for the DL ensemble model (Figure 4.13-B).



A. ML Model



B. DL Model

Figure 4.13. Combined Violin plot of the true and predicted labels of the ML/DL ensemble model

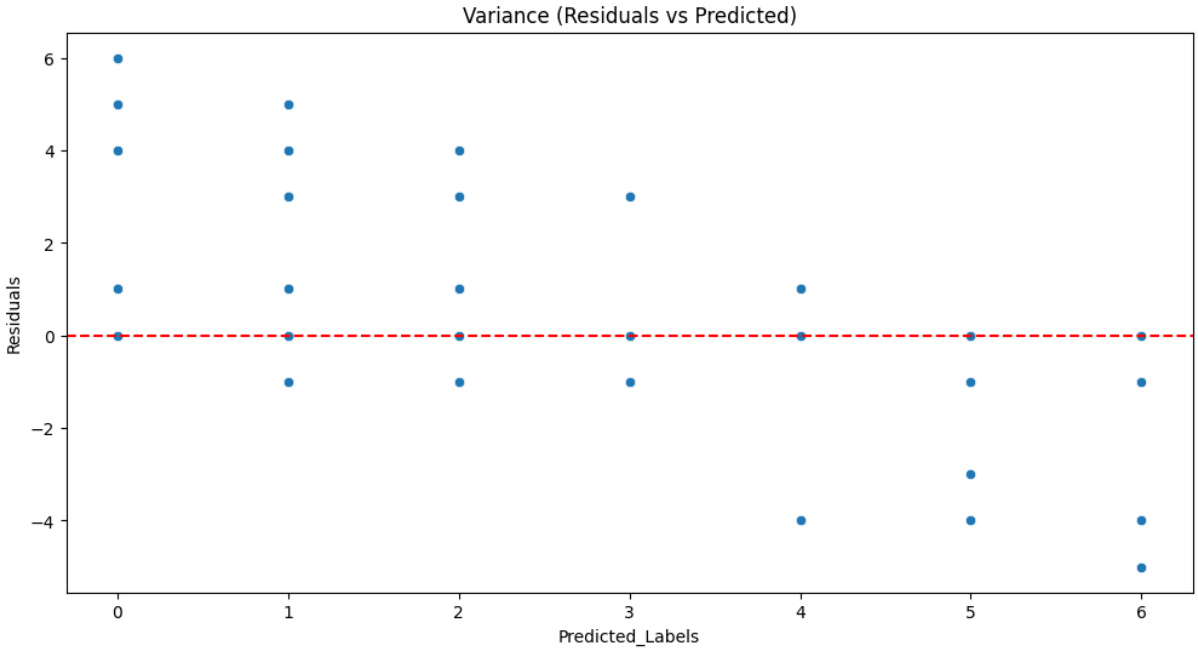
Violin plot is one of the performance evaluation charts that can be used to assess the performance of the trained ML and DL models using a mix of box plot and kernel density distribution. The Violin chart in our study is constructed using the predicted and true labels of the cancer scores. As seen in Figure 4.13, the distribution of the true and predicted labels almost the same. This proves the high accuracy of the proposed model in predicting the true labels with low number of errors. In this branch of study, we utilized the violin chart in order to judge the classification problem more precisely. In the fourth branch of the study, the adjacent classes are merged to transfer the problem into a specific classification task so the Violin chart can help to assess the classification task especially in case of multi-class classification (Fourth branch of the study is considered as multi-class classification).

4.5.1.2. Variance Analysis

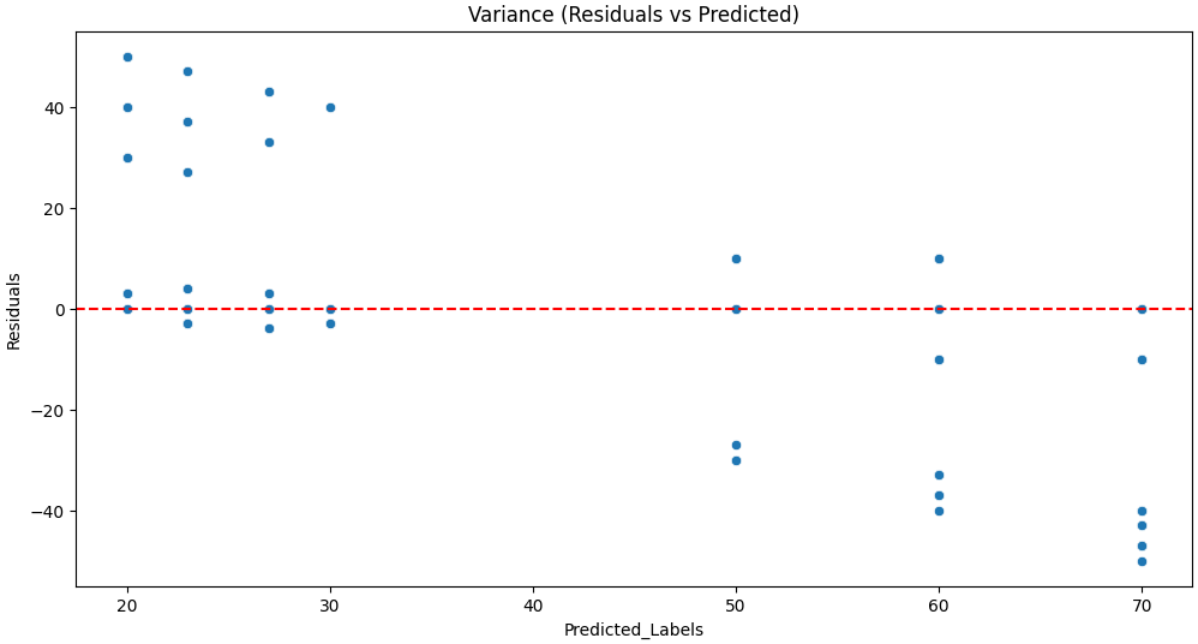
In this part, we will analyze the variance of our predictions and the original breast cancer risk scores. Figure 4.14 (on both cases of ML and DL ensemble models) shows that most of variances are located near 0 value. The obtained variances of both ML and DL ensemble models are [-40, +60] and [-40, +40] and these variances are due to the problem of the unbalanced dataset (which can't be totally solved by the oversampling operation since the test set can't be oversampled). The number of variances with wide range are 10 and 12 for both ML and DL ensemble models, respectively. While the number of variances with low range are 13 and 16 for both ML and DL, respectively.

However, although the variance range is not small, but the number of error samples (caused the variance problem) is too small comparing to the total number of test samples.

Figure 4.14 includes the variance results of both ML and DL models.



A. ML ensemble model



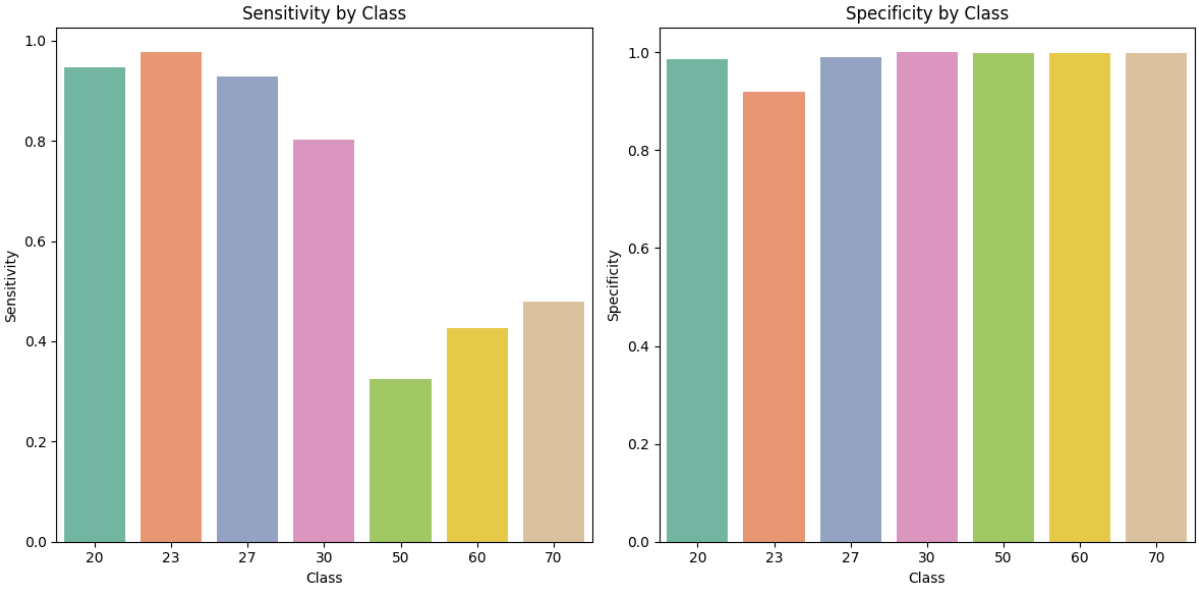
B. DL ensemble model

Figure 4.14. Variance plot of the true and predicted labels of the ML ensemble model

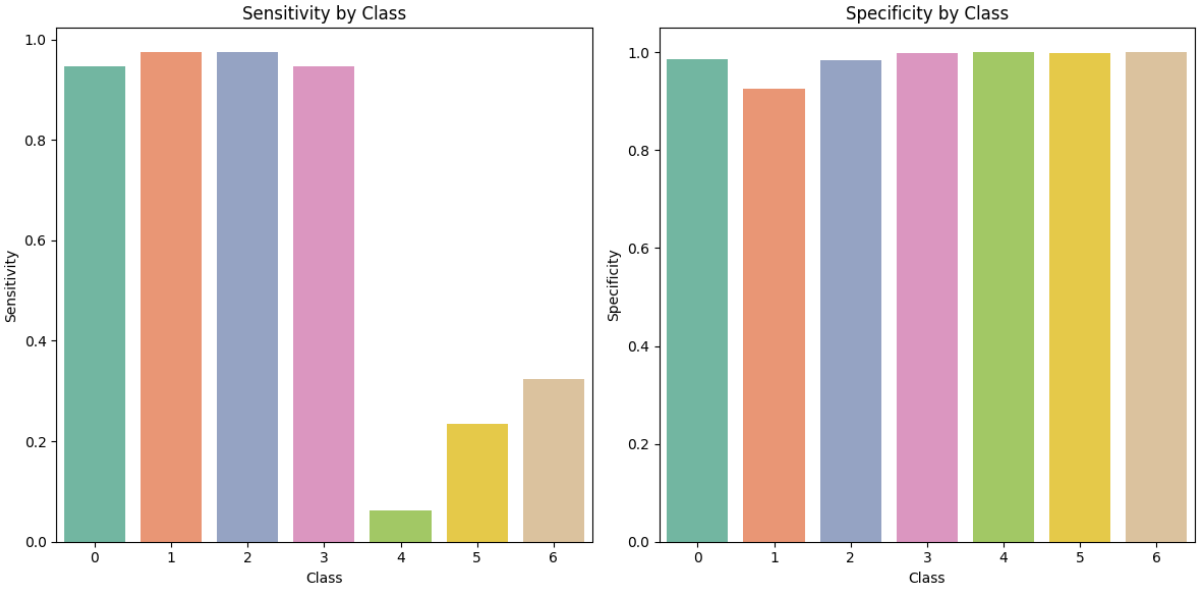
4.5.1.3. Sensitivity and specificity results

In many medical decision support systems, the sensitivity and specificity test are performed to ensure the accuracy and robustness of the proposed models.

Figure 4.15-A and B include the sensitivity and specificity of the ML and DL ensemble models.



A. ML Ensemble Model



B. DL Ensemble Model

Figure 4.15. Sensitivity and Specificity plot of the true and predicted labels of the ML/DL ensemble model

Figure 4.15 proves that the specificity of all cancer scores (all categories) are high in both ML and DL ensemble models.

For sensitivity, the low and medium cancer score categories have high values. However, the high-risk score categories 50-70 registered low sensitivity values due to the fact that the number of their test samples are too small comparing to the number of samples of the other categories (which is mentioned earlier in this branch of study).

4.6. Results of the fifth branch (Regression model)

The final thread of this study will be performed using regression models. We will treat the target column as a continuous score value, so no merging operations is applied here. In this part, the dataset obtained from the fourth branch is utilized.

First, the target column is transformed using the logarithm transform to minimize the wide range of the target column and enhance the performance of the regression model.

Figure 4.16 shows the target column (breast cancer prediction) before and after transform.

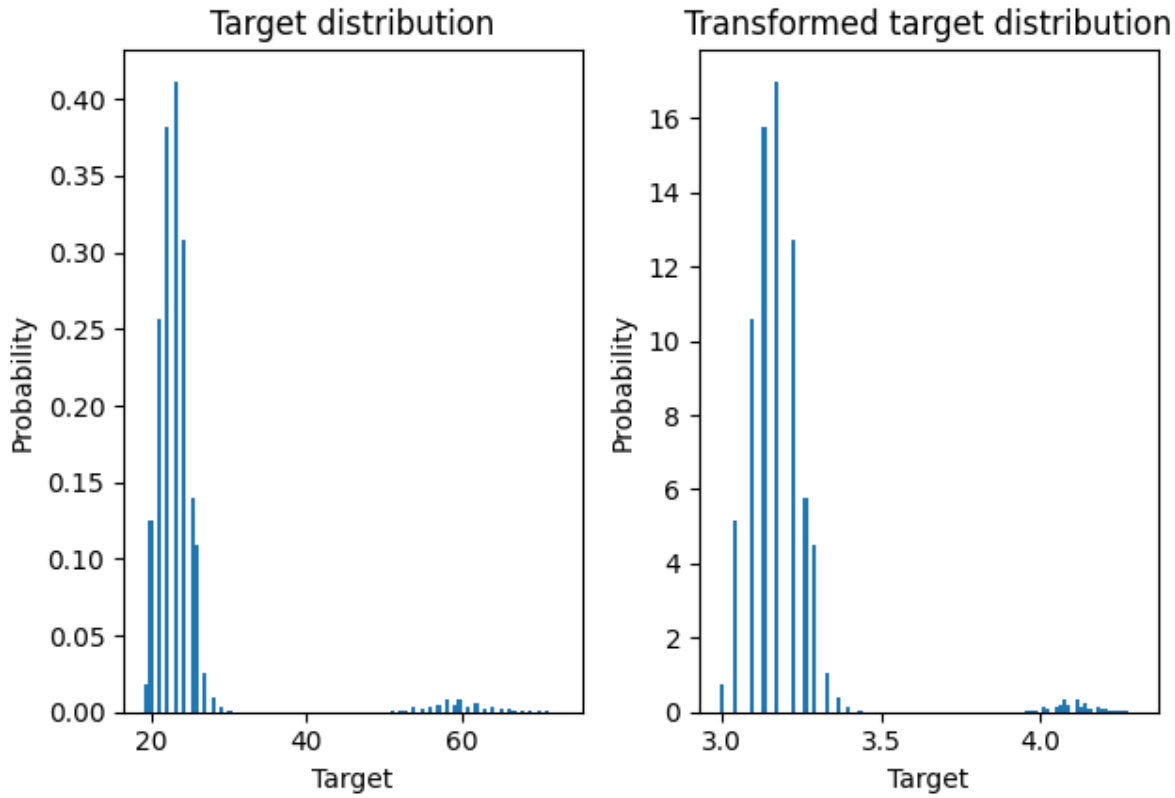


Figure 4.16. Target column before and after applying the logarithm transform

After that, three regression models are applied: the decision trees regression DTR, the random forest regression RFR, and the K-Nearest neighbor regression model. Then, an ensemble of these three regression models is constructed and evaluated. The dataset is also split into 80% training and 20% test as in the previous scenario.

Table 4.10 includes the evaluation results of the regression models using the regression metrics (MSE, MedAE).

Table 4.8 Evaluation results of the trained ML and DL models of the fifth scenario

Model	**MSE**	**MedAE**
RFR	0.0164	0.019
KNN Regression	0.03125	0.023
DTR	0.029	0
Ensemble ML	0.0104	0
DL model	0.11	0.0205

The best regression model is the ML ensemble model with 0.0104 mean squared error. The low value of the regression model proves his ability to predict the range-based breast cancer risk value in a high accuracy.

Figure 4.17 shows the distribution of the actual and predicted risk score using the ML ensemble regression model.

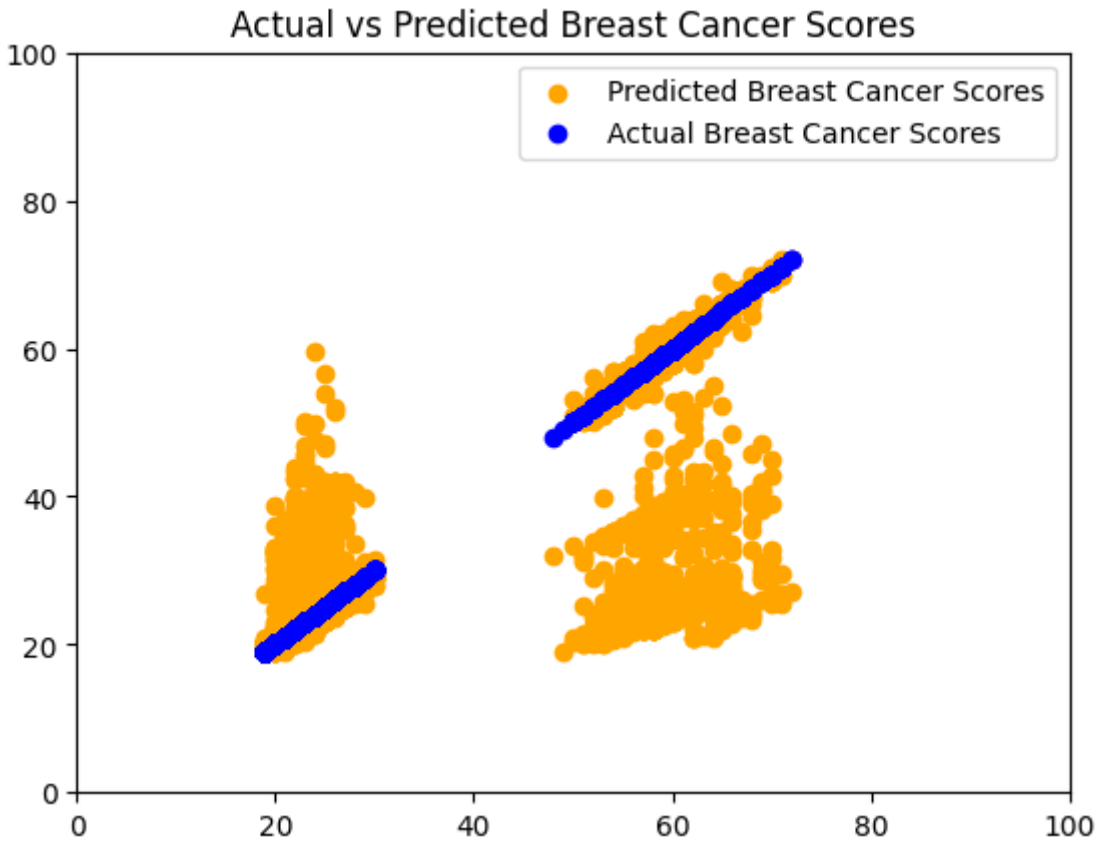
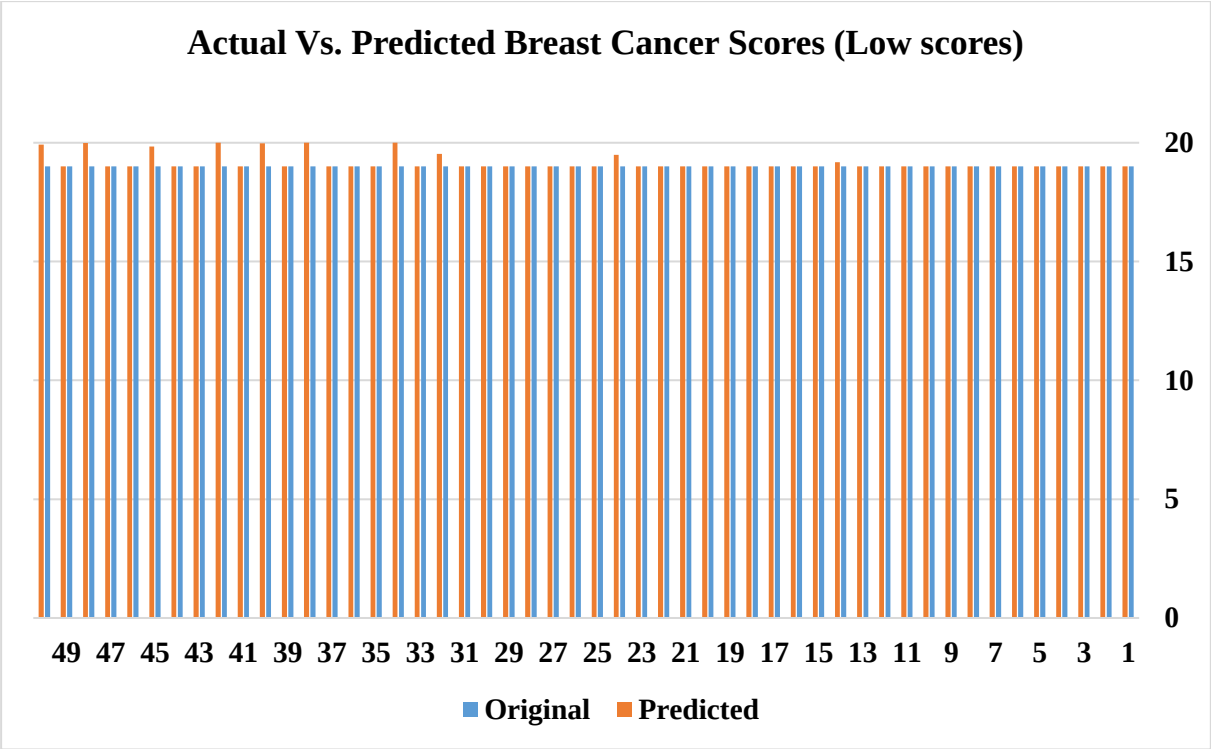
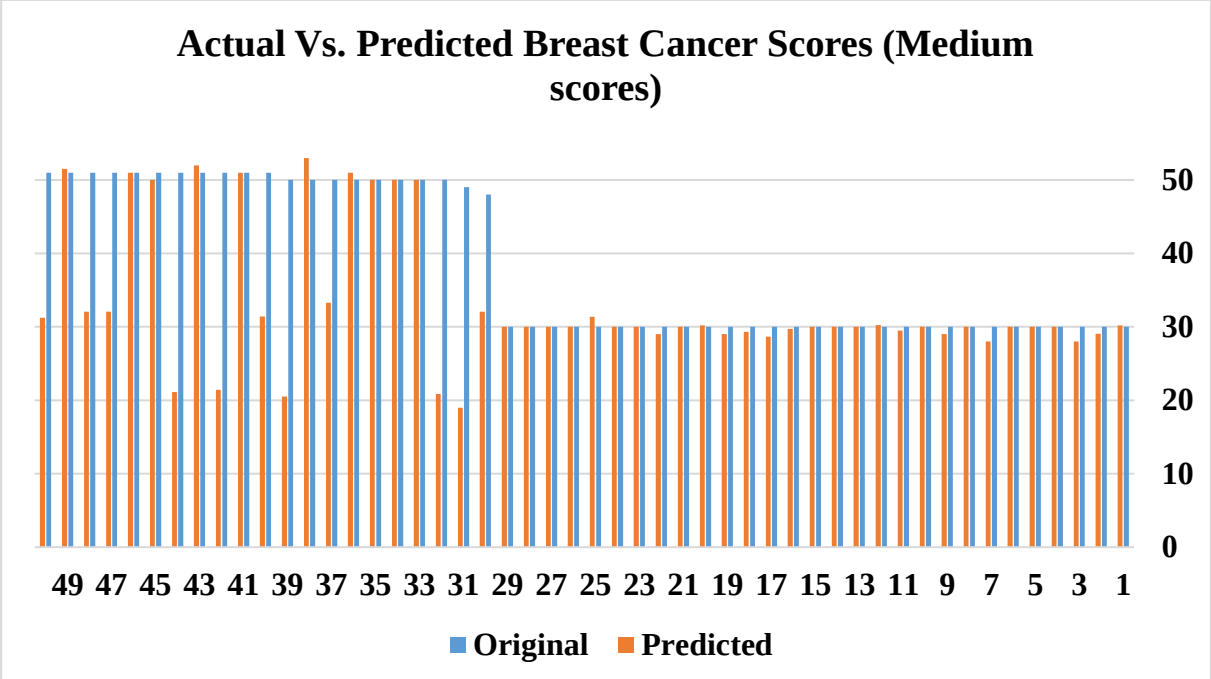


Figure 4.17. The actual and predicted breast cancer score according to the ML ensemble regression model.

Figure 4.17 demonstrates two clusters. One of them in the low-risk range, while another one in the high-risk range. However, the predicted values are distributed around each cluster with a variance which is notable especially in case of high-risk score samples. This result matches the result of the fourth branch and again it can be interpreted due to the distribution nature of the test samples which are extremely biased to the low-range cancer score.

Now for more accurate judgment, we extract three different parts of the distribution plot of the actual and predicted scores at different score ranges. The first one is in the low-range score samples, the second one is in the medium-range score, while the third one is intended for the high-range score. Figure 4.18 illustrates these three comparisons between the actual and predicted risk score of the three ranges.


Actual Vs. Predicted Breast Cancer Scores (Low scores)
20
15
10
5
0
49 47 45 43 41 39 37 35 33 31 29 27 25 23 21 19 17 15 13 11 9 7 5 3 1
Original
Predicted


Actual Vs. Predicted Breast Cancer Scores (Medium scores)
50
40
30
20
10
0
49 47 45 43 41 39 37 35 33 31 29 27 25 23 21 19 17 15 13 11 9 7 5 3 1
Original
Predicted

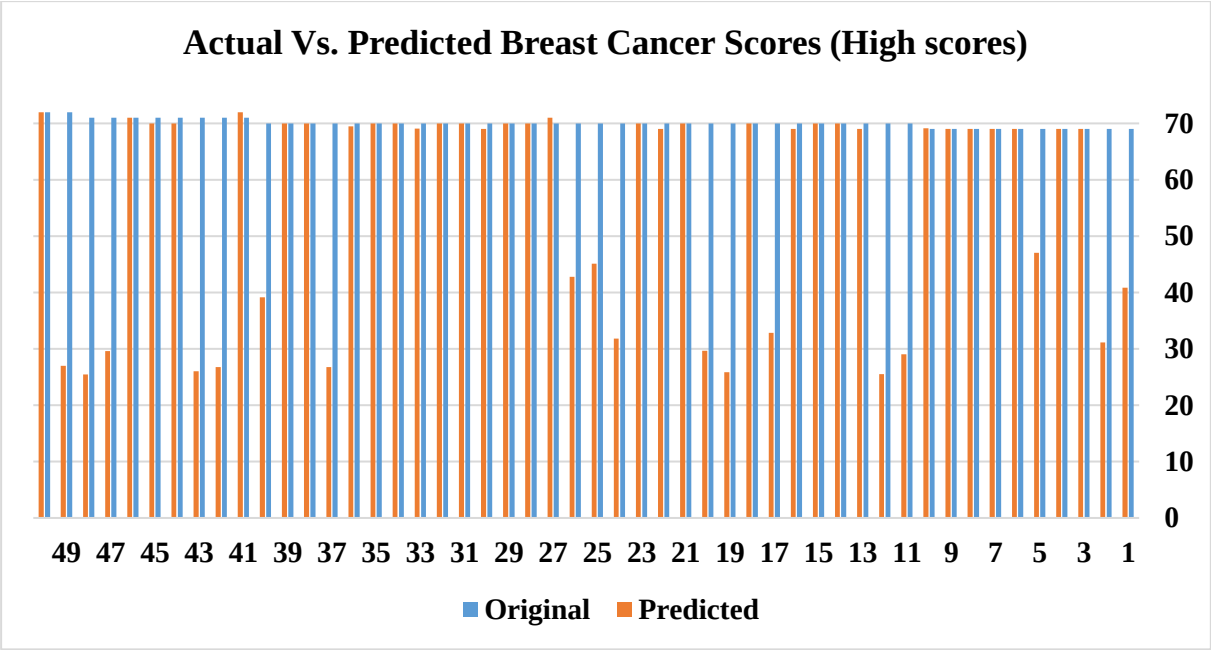


Figure 4.18. Actual and predicted risk score of the three ranges

For the low score, the match between the predicted and original scores are too high. The medium range includes a few variances, while the most variances can be noticed in the high range.

However, although there are some variances in the prediction of the high score levels, but the prediction still produces accurate and closed risk score.

We also need to clarify that the regression method has its limit in performance (as seen in Figures 4-17 and 4-18). The target column in this case represents a wide range of possible values of cancer predictions. So, any little change in the predicted score will result in wrong prediction or biased prediction. However, future studies can focus on improving the performance of such regression-based cancer prediction methods.

Chapter 5

Conclusion and future work

5.1. Comparison with the related state-of-art

Table 5.1 includes a comparison between the current study and the previous state-of-art in order to define my research importance and contribution.

Table 5.1. Comparison between the current study and related work.

Researcher	Outcome	Results / Limitations
My study	Range-Based score	Accuracy: 95.8%. Fixed cancer score.
	Range-Based score	Original Accuracy: 85.3%, ±1 variation: 91.12%, ±2 Variation 91.33%.
	Range-Based score	Accuracy 99.98%.
Hussain et al. [58]	Cancer prediction (Yes: 1, No: 0)	GoogleNet: 99.26% accuracy, AlexNet: 99.26% accuracy. Small dataset.
Guo et al. [23]	Cancer prediction (Yes: 1, No: 0)	Accuracy: 98.79%. Binary prediction limitation.
Uddin et al. [49]	Cancer prediction (Yes: 1, No: 0)	Voting classifier: 98.77% accuracy. Small dataset.
Li and Sundararajan [44]	Cancer prediction (Yes: 1, No: 0)	SVM accuracy: 96.6%, Bayes accuracy: 91.26%. Limited dataset size.
Leventi et al. [57]	Cancer prediction (Yes: 1, No: 0)	Accuracy: 95.32%. Small dataset, low accuracy for breast cancer prediction.
Kayikci et al. [59]	Cancer prediction (Yes: 1, No: 0)	AUC: 0.95, accuracy: 91.2%, precision: 84.1%, recall: 79.8%. Small dataset.
Saleh et al. [48]	Cancer prediction (Yes: 1, No: 0)	Accuracy: 95.18%. Small dataset.
Kurian et al. [53]	Cancer prediction (Yes: 1, No: 0)	Decision tree achieved highest accuracy (94.30%). Binary prediction limitation.
Ashokkumar et al. [47]	Axillary Lymph Node Metastasis prediction (Yes: 1, No: 0)	Accuracy: 94%. Limited dataset size.
Ming et al. [45]	Three cases (High, moderate, low risk)	Accuracy: 84.3% - 88.9%. Need more risk factors.
Botlagunta et al. [50]	Cancer prediction (Yes: 1, No: 0)	DT classifier: 83% accuracy, 0.87 AUC. Small dataset.
Rajendran [34]	Cancer prediction (Yes: 1, No: 0)	Best accuracy: 99.1%, low sensitivity (78.1%). Small dataset.

Yang et al. [52]	Cancer Recurrence prediction (Yes: 1, No: 0)	Training accuracy: ANN: 73.55%, Markov Model: 76.07%, Fusion: 75.63%. False positives and negatives.
Al-Jawad et al. [18]	Survival status (1 or 2)	SVM recall: 73.78%, Precision: 74.77%, BN recall: 78.22%, Precision: 64.47%. Small dataset.
Hou et al. [33]	Cancer prediction (Yes: 1, No: 0)	DNN and RF accuracy: 72.8%, XGBoost accuracy: 74.2%. Small dataset.

In contrast to previous studies, my research utilizes a significantly larger dataset (BCSC with 317880 samples) and applies a fusion of machine learning and deep learning techniques, which yield an exceptionally high accuracy rate of 99.98%. Moreover, the novelty of a range-based cancer score system and hyperparameters optimization in my study addresses some of the limitations found in other research, providing a more nuanced and efficient predictive model for breast cancer.

5.2. Conclusion

This study offers novel insights into breast cancer prediction, leveraging machine learning and deep learning techniques to build a robust, predictive model. The research bifurcated into five primary sections: risk factor weighting, range-based cancer prediction, and a fusion model for prediction. The first section presented a weighting algorithm applied to the BCSC dataset, improving accuracy to 95.8% and minimizing errors. The second part introduced a range-based predictive model, utilizing a probabilistic model to augment the BCSC dataset. Bayesian hyperparameters optimization was used for training the ensemble learning model, yielding superior True Positive Rate (TPR), Positive Predictive Rate (PPR), and accuracy in scenarios allowing for class-variance tolerance. Furthermore, the enhanced BCSC dataset provides a more granular understanding of cancer prediction. The third section proposed a fusion-based prediction system, combining an LSTM deep learning model with an ensemble of boosted decision trees, leading to an improved accuracy of 99.98% when using the fusion approach. This research not only offers innovative methodologies in breast cancer prediction but also provides an enriched version of the BCSC dataset for future investigations. The integration of machine learning and deep learning techniques, coupled with a comprehensive understanding of breast cancer prediction, will significantly contribute to the field of medical engineering.

The experiments were also applied using the individual modes and the fusion approach to measure the effect of the fusion approach on performance. The results demonstrated an improvement in performance using the fusion approach so that the accuracy was 99.98%.

The forth part of this study concentrates on rebuilding a new breast cancer dataset starting from the original (unbalanced) dataset by applying the probabilistic model to transform the target column into a range-based one. Many ML and DL with ensemble learning models are applied in this branch, and the results are evaluated in terms of many statistical and medical analyses.

For the final branch of the study, the same dataset of the fourth branch is utilized but the target column is considered as a continuous column, so many regression models are applied with ensemble regression model and the results are also analyzed and discussed. Although the results of the classification ensemble models were efficient and promised, the regression-based method had its limits in performance due to the nature of the wide range of target column (cancer prediction score).

5.3. New scientifc contributions

The main scientifc contributions of the dissertation are summarized in the following theses.

Thesis I.

I introduced an innovative breast cancer prediction model based on a sophisticated weighting algorithm applied to the BCSC dataset. The model encompasses a multi-step process, starting with dataset normalization and balancing, followed by a novel weighting algorithm incorporating expert opinions and international medical reports. The final degree of importance (DOI) is determined, influencing suggested training weights for risk factors. The optimization tree model is selected for its adaptability to hyperparameters and handling of data complexities. Empirical results demonstrate a 6.9% performance improvement, with substantial reductions in False Discovery and False Negative Rates. Notably, risk factor analyses identify "Race" as the most influential, underscoring its critical role in predictive accuracy.

Thesis II.

I proposed a novel Range-based breast cancer prediction model, an extension of Thesis I, comprising two integral systems: breast-cancer factors weighting and a statistical model for computing essential breast cancer statistics. The mathematical model calculates the range-based

cancer prediction score using Bayes' theorem, incorporating suggested training weights and risk factor probabilities. This model is employed to create new subclasses within the BCSC dataset, introducing three attributes: cancer score, non-cancer score, and final prediction. Machine learning training is conducted using modified dataset versions, considering two scenarios: a subset of BCSC and the entire dataset. The probabilistic model is applied to evaluate and compute final prediction scores, leading to a new distribution of result prediction scores, with subclasses for low and high-predicted percentages of breast cancer. I proved that my range-based model achieved average TPR values of 94.61% and 90.15% for both sub and entire datasets, respectively. The average PPR values of the sub and entire datasets are 95.28% and 85.55%, respectively. I also applied experiments using ±1 and ±2 class variance (Classes "19", "20" and "21" for example is considered as one category). The total 36 classes are concluded into only 7 categories. I showed that the accuracy is increased by 5.82% and 6.03% for ±1 and ±2 class-variances, respectively.

Thesis III:

Utilizing the range-based and balanced BCSC dataset from the previous parts, I introduced a novel Range-based breast cancer prediction approach employing a fused DL-ML model. The initial step involved categorizing classes into seven categories through a "Grouping step," resulting in a new BCSC dataset enriched with added knowledge. The dataset was then split into training and test sets for the development of both a deep learning (DL) architecture (LSTM and Dense layers) and an ensemble learning model. In the final step, a score-level fusion technique was applied to combine the ML and DL models, enhancing overall performance.

Multiple experimental scenarios were executed to assess the proposed method, incorporating modifications to the LSTM architecture, changes in the number of neurons, learning epochs, and variations in the training and test percentages. The results demonstrated superior performance of the fused model compared to individual ML and DL models, with an accuracy increase of 1.08% and 3.3%, TPR improvement by 1.66% and 5.46%, and PPR enhancement by 2.01% and 5.44% compared to DL and ML individual models. These findings affirm the significant performance improvement achieved through the fusion of DL and ML models.

Sub-Thesis I:

I proposed a novel Classification-based range-based ensemble model for the original BCSC dataset, employing this probabilistic model to compute a new distribution of the target column. A detailed exploration ensued, introducing two ensemble approaches: a Machine Learning ensemble featuring Decision Trees (DT), Random Forest (RF), and Light Gradient Boosting Machine (LGMB), and a Deep Learning ensemble incorporating Long Short-Term Memory (LSTM) and 1D-Convolutional Neural Network (1D-CNN). Through rigorous analyses encompassing violin distribution examination and variance analysis, the study offers insights into model accuracy and the impact of class imbalances on predictions. Notably, the results demonstrate the model's high accuracy in predicting breast cancer categories, evident in the close alignment of the original and predicted cancer risk distributions. Additionally, the section addresses the nuanced metrics of sensitivity and specificity in medical decision support systems, particularly focusing on challenges posed by smaller sample sizes, especially in high-risk categories.

Sub-Thesis II:

I proposed a regression-based and range-based breast cancer model. I directed my efforts towards the utilization of regression analysis to predict continuous breast cancer risk scores, as the new range-based score represents a continuous scope. The dataset, obtained from the fourth branch of the study, underwent a logarithmic transformation on the target column. This transformation was instrumental in normalizing the target's distribution, thereby enhancing the predictive efficacy of the regression models. I employed three distinct regression models—Decision Tree Regression (DTR), Random Forest Regression (RFR), and K-Nearest Neighbor (KNN) Regression—followed by the construction of an ensemble model aggregating these three. The evaluation of regression breast cancer models was performed using regression-specific metrics such as Mean Squared Error (MSE) and Median Absolute Error (MedAE). The ensemble model exhibited remarkable precision, recording the lowest MSE among the cohort, substantiating its refined predictive capability. Despite the observed variance in high-risk score predictions, the model's output remains closely aligned with the actual risk scores, underlining the robustness and accuracy of the regression approach employed in this study.

5.4. Recommendation and Future Work

Future studies can get benefit of the modified version of the BCSC dataset and conduct more experiments.

Future research can focus on utilizing other breast cancer datasets and studying the effect of incorporating different risk factors on the performance of range-based breast cancer prediction. Other possible experiments can be conducted like designing a multimodal breast cancer prediction model using x-ray or CT-scan images along with the risk factors improving the reliability of the breast cancer prediction models.

Acknowledgement

Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C), available at: http://www.bcsc-research.org/.

This article publication was funded by PPCU supported by NKFIH, financed under Thematic Excellence Programme (TUDFO/51757-1/2019-ITM).

The publications and part of the academic guidance was achieved by the co-advisor Dr. Ali Mahmoud Mayya who is specialist in machine learning and deep learning and has published many researches in scientific journals indexed in Scopus and Web of Science.

Appendices

Appendix A- Breast Cancer Risk Factors Evaluation

Breast Cancer Risk Factors Evaluation

This form is designed in order to analysis the known risk factors of breast cancer and determine the impact of each one on the future possible risk. Please enter your name, country and specialty. You only need to choose one of three values (low, medium, high) for each risk factor to indicate the impact of this factor in the final risk degree of breast cancer. Note: These risk factors are parts of the "Breast Cancer Surveillance Consortium" database. For more details you can visit: https://www.bcsc-research.org/data/rfdataset/dataset.
If you are not in the same field of this topic, please ignore this form.

* Indicates required question

1. Full name

_____

2. Country *

_____

3. Specialty *

_____

4. Menopause

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

5. Age (as groups)

Mark only one oval.

- Low
- Medium
- High

6. Breast density

Mark only one oval.

- Low
- Medium
- High

7. Race

Mark only one oval.

- Low
- Medium
- High

8. Hispanic Originality

Mark only one oval.

- Low
- Medium
- High

9. Body mass index

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

10. Age at first birth

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

11. Number of first degree relatives with breast cancer

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

12. Previous breast procedure

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

13. Result of last mammogram before the index mammogram

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

14. Surgical menopause

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

15. Current hormone therapy

Mark only one oval.

- ◯ Low
- ◯ Medium
- ◯ High

Appendix B- Three trials of the results of the first branch of the study

Three trials of the trained ensemble model of the first part of the study are shown in Table B-1.

Table B-1. Evaluation of the risk estimation model using the weighted and non-weighted version of the risk factors (Three different trials).

	With Weighting (%)	Without Weighting (%)
Majority class FNR (Trial1)	3.3	8.3
Majority class FNR (Trial2)	3.3	8.3
Majority class FNR (Trial3)	3.28	8.27
Minor class FNR (Trial1)	10.5	28.1
Minor class FNR(Trial2)	10.5	28.1
Minor class FNR (Trial3)	11.1	29.02
Majority class FDR (Trial1)	1.8	5
Majority class FDR (Trial2)	1.81	5.1
Majority class FDR (Trial3)	1.82	5.3
Minor class FDR (Trial1)	17.5	40.1
Minor class FDR (Trial2)	17.6	40.2
Minor class FDR (Trial3)	17.5	40.2
Overall Validation Accuracy (Trial1)	95.7	88.8
Overall Validation Accuracy (Trial2)	95.6	88.7
Overall Validation Accuracy (Trial3)	95.85	88.9
Training Time (Trial1)	38.65	41.09
Training Time (Trial2)	40.50	41.20
Training Time (Trial3)	40.32	41.25

Appendix C- Three trials of the results of the second branch of the study

Three trials of the entire dataset results (Second part of the study)

Original Results Trial 1: Accuracy=85.3%, Trial 2: Accuracy=85.1%, Trial 3: Accuracy=85.5%

	TPR (Trial1)	TPR (Trial2)	TPR (Trial3)	FNR (Trial1)	FNR (Trial2)	FNR (Trial3)	PPR (Trial1)	PPR (Trial2)	PPR (Trial3)	FDR (Trial1)	FDR (Trial 2)	FDR (Trial 3)
19	77.9	77.8	77.8	22.1	22.2	77.9	86.7	86.6	86.7	13.3	13.4	13.3
20	83.9	84.1	84.1	16.1	15.9	83.9	87.5	87.6	87.7	12.5	12.4	12.3
21	84.4	84.4	84.4	16.6	15.6	83.4	85.3	85.3	85.3	14.7	14.7	14.7
22	84.4	84.4	84.4	16.6	15.6	83.4	85.6	85.5	85.7	14.4	14.5	14.3
23	83.9	83.7	83.7	16.1	16.3	83.9	85.9	85.9	85.9	14.1	14.1	14.1
24	82.9	82.9	82.9	17.1	17.1	82.9	85.2	85.2	85.2	14.8	14.8	14.8
25	75.3	75.2	75.5	24.7	24.5	75.3	80.6	80.6	80.6	19.4	19.4	19.4
26	83.9	83.8	84.1	16.1	15.9	83.9	85.3	85.3	85.4	14.7	14.7	14.6
27	72.3	72.3	72.5	27.7	27.5	72.3	78.6	78.6	78.6	21.4	21.4	21.4
28	64.3	64.2	64.4	35.7	35.6	64.3	72.3	72.3	72.4	27.7	27.7	27.6
29	64.5	64.5	64.5	35.5	35.5	64.5	69.6	70	70	30.4	30	30
30	69.4	69.4	69.4	30.6	30.6	69.4	73.5	73.5	73.5	26.5	26.5	26.5
31	0	0	0	100	100	0	-	-	-	100	#VALUE!	#VALUE!
48	100	100	100	0	0	100	100	100	100	0	0	0
49	100	100	100	0	0	100	94.1	94.1	94.3	5.9	5.9	5.7
50	100	100	100	0	0	100	82.1	82	82.1	17.9	18	17.9
51	100	100	100	0	0	100	86	86	86	14.0	14	14
52	100	100	100	0	0	100	81.9	81.9	82.1	18.1	18.1	17.9
53	100	100	100	0	0	100	83.6	83.6	83.6	16.4	16.4	16.4
54	96.6	96.6	96.7	3.4	3.3	96.6	83.1	83.1	83.1	16.9	16.9	16.9

55	99.2	99.2	99.3	0.8	0.7	99.2	84.0	84.0	84.2	16.0	16	15.8
56	100	100	100	0	0	100	85.8	85.8	86	14.2	14.2	14
57	99.6	99.6	99.6	1.4	0.4	98.6	86.8	86.8	86.8	13.2	13.2	13.2
58	99.2	99.2	99.2	0.8	0.8	99.2	86.4	86.2	86.4	13.6	13.8	13.6
59	100	100	100	0	0	100	84.9	84.9	85.3	15.1	15.1	14.7
60	98.6	98.6	98.8	1.4	1.2	98.6	84.8	84.8	84.8	15.2	15.2	15.2
61	99.3	99.3	99.5	0.7	0.5	99.3	87.2	87.2	87.2	12.8	12.8	12.8
62	100	100	100	0	0	100	87.1	87.1	87.3	12.9	12.9	12.7
63	99.1	99.1	99	0.9	1	99.1	85.3	85.3	85.3	14.7	14.7	14.7
64	99.3	99.3	99.4	0.7	0.6	99.3	85.8	85.8	85.8	14.2	14.2	14.2
65	98.1	98.1	98.3	1.9	1.7	98.1	89.8	89.8	90	10.2	10.2	10
66	100	100	100	0	0	100	89.5	89.1	89.5	10.5	10.9	10.5
67	100	100	100	0	0	100	83.4	83.4	83.4	16.6	16.6	16.6
68	100	100	100	0	0	100	93.0	93.0	93.0	7.0	7	7
69	100	100	100	0	0	100	90.1	90.1	90.1	9.9	9.9	9.9
70	100	100	100	0	0	100	90.4	90.4	90.4	9.6	9.6	9.6
71	100	100	100	0	0	100	90.5	90.5	90.5	9.5	9.5	9.5
72	100	100	100	0	0	100	79.2	79.2	79.5	20.8	20.8	20.5
73	100	100	100	0	0	100	100	100	100	0	0	0

Figures C-1 and C-2 also shows the confusion matrix of two different number of epochs (the second branch of the study). Figure C-3 also illustrates the training curves using 15 and 20 training epochs.

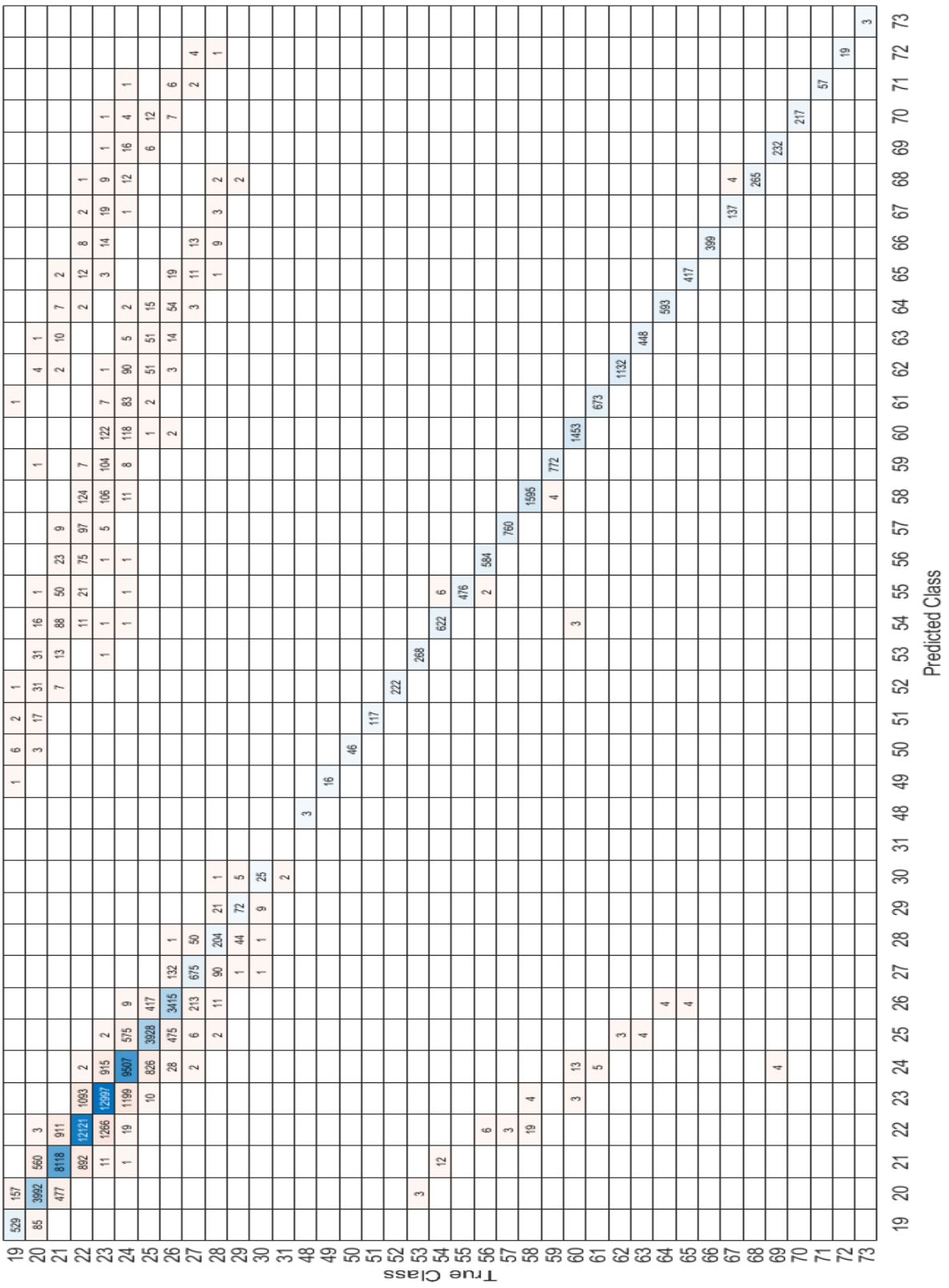


Figure C-1 Confusion matrix using 15 training epochs (the second branch of the study)

Figure C-2 Confusion matrix using 20 training epochs (the second branch of the study)

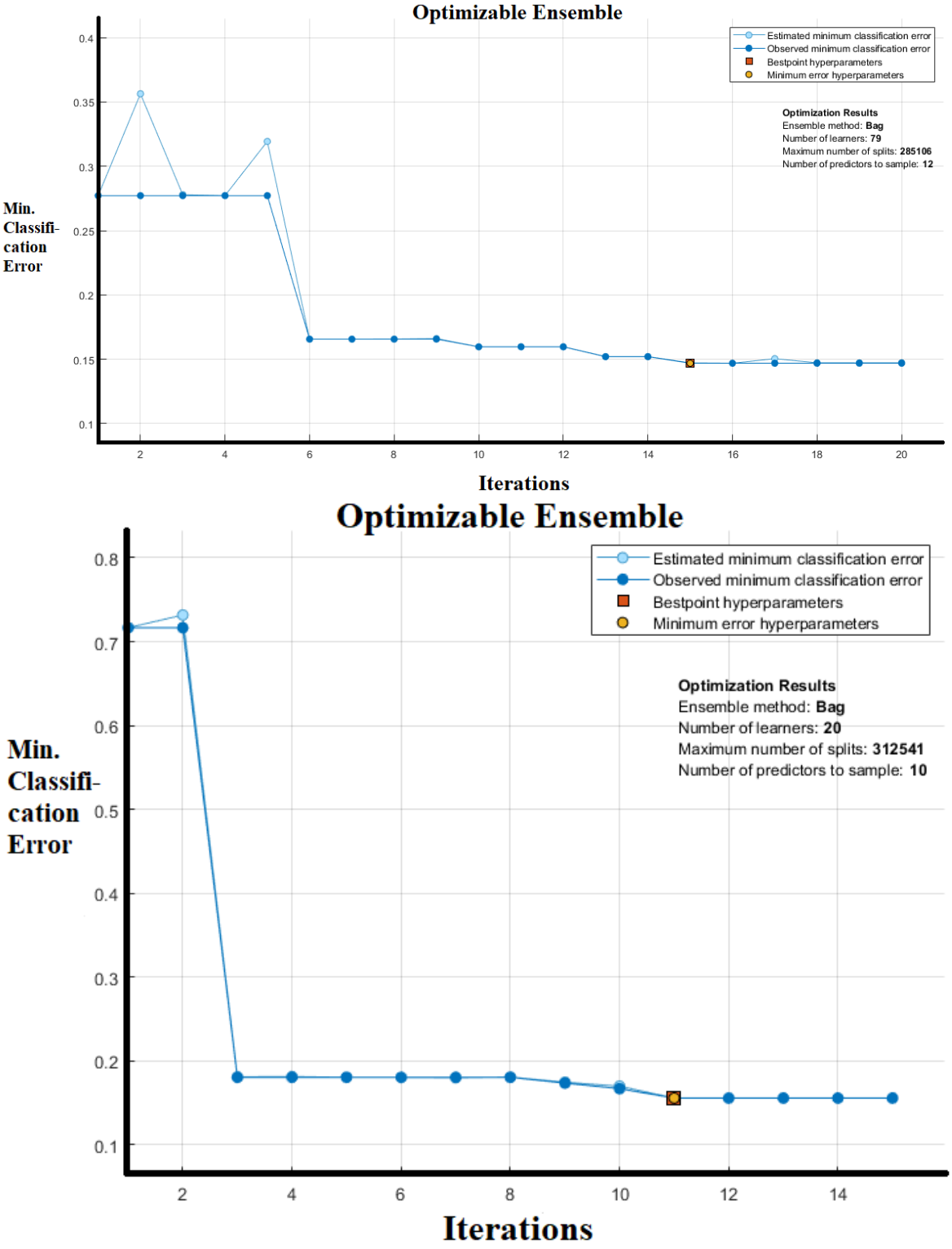


Figure C-3 Training curves of the optimizable ensemble model using 15 and 20 training epochs.

Appendix D- Three trials of the results of the fourth branch of the study

Table D-1 shows three trials of the same results of training ML and DL models of the fourth branch of our study.

Model	Accuracy %	Precision %	Recall %	F1-score %
LGBM	94.19	94	94	94
DT	90.83	91.13	90.83	90.97
RF	93.35	93.09	93.35	92.97
Ensemble (LGBM, DT, RF)	94.35	94.1	94.35	94.03
1D-CNN	93.12	90.46	93.12	91.7
LSTM	93.63	93.22	93.63	93.19
Ensemble of 1D-CNN and LSTM	94.35	94.1	94.35	93.62

Table D-1 Three Trails of evaluation results of the trained ML and DL models of the fourth scenario

Model	Accuracy % Trial1	Accuracy % Trial2	Accuracy % Trial3	Precision % Trial1	Precision % Trial2	Precision % Trial3	Recall % Trial1	Recall % Trial2	Recall % Trial3	F1-score % Trial1	F1-score % Trial2	F1-score % Trial3
LGBM	94.19	94.22	93.99	94	94	94	94	93.8	93.9	94	93.89	93.94
DT	90.83	90.75	90.90	91.13	91.22	91	90.83	90.8	90.8	90.97	91	90.89
RF	93.35	94.2	93.7	93.09	93.2	93.1	93.35	94	94.1	92.97	93.59	93.59
Ensemble (LGB	94.35	94.5	94.2	94.1	94.2	94.2	94.35	94.7	94.5	94.03	94.4	94.34

M, DT, RF)												
1D-CNN	93.12	93.25	93.2	90.46	90.35	90.4	93.12	93.2	93.3	91.7	91.75	91.82
LSTM	93.63	94	94	93.22	93.6	93.5	93.63	93.5	93.6	93.19	93.54	93.54
Ensemble of 1D-CNN and LSTM	94.35	94.5	94.3	94.1	94	94	94.35	93.5	93.5	93.62	93.74	93.74

Appendix E- Three trials of the results of the fourth branch of the study

Three trials are also performed in the last branch of this study. The results of these trials are shown in Table E.1 bellow.

Table E.1 Three trials of the evaluation results of the trained ML and DL models of the fifth scenario

Model	MSE Trial1	MSE Trial2	MSE Trial3	MedAE Trial1	MedAE Trial2	MedAE Trial3
RFR	0.0164	0.0163	0.01644	0.019	0.01889	0.0189
KNN Regression	0.03125	0.0312	0.0325	0.023	0.023	0.0229
DTR	0.029	0.02877	0.0289	0	0	0
Ensemble ML	0.0104	0.0103	0.0103	0	0	0
DL model	0.11	0.114	0.109	0.0205	0.0089	0.009

Appendix F- Hyperparameters optimization details

Table F-1. Hyperparameters in our experiments (Ensemble models).

Hyperparameter	Initial Value
Maximum Number of Splits (range)	[1, max(2, n–1)], where n is the number of observations
Ensemble Method	AdaBoost (Boosting methods)
Minimum leaf size (per decision tree)	8
Number of learners (Decision Trees)	30
Initial Learning rate	0.1
Optimizer	Bayesian Optimization
Iterations	15/20
Maximum training time	300

At the end of optimization, we got different values of the optimizable hyperparameters which were clearly different form the initial values. For example, number of learners was 73, the final learning rate was 0.8, the maximum number of splits was 6113.

References

[1] M. Fatih, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare,* vol. 8, no. 2, p. 111, 2020.

[2] S. Patil, I. H. Moafa, M. M. Alfaifi, A. M. Abdu, M. A. Jafer, L. Raju, A. T. Raj and S. M. Sait, "Reviewing the role of artificial intelligence in cancer," *Asian Pacific Journal of Cancer Biology,* vol. 5, no. 4, pp. 189-199, 2020.

[3] V. K. Kamal and D. Kumari, "Use of artificial intelligence/machine learning in cancer research during the COVID-19 pandemic," *Asian Pacific Journal of Cancer Care,* vol. 5, no. S1, pp. 251-253, 2020.

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal ,* vol. 13, pp. 8-17, 2015.

[5] G. A. Colditz and E. K. Wei, "Risk prediction models: applications in cancer prevention," *Current Epidemiology Reports,* vol. 2, pp. 245-250, 2015.

[6] A. S. Ahmad and A. M. Mayya, "A new tool to predict lung cancer based on risk factors," *Heliyon ,* vol. 6, no. 2, 2020.

[7] Z. LAKY, "Cancer prevention: Modifiable risk factors," in *ENVI Workshop Proceedings*, 2020.

[8] American Cancer Society, "Breast Cancer Facts & Figures 2019-2020," American Cancer Society, Atlanta, 2019.

[9] American Cancer Society, "Breast Cancer Risk and Prevention," American Cancer Society, Atlanta, 2019.

[10] American Cancer Society, "Breast Cancer Fact Sheet," American Cancer Society, Atlanta , 2020.

[11] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen and S. S. Iyengar, "Computational health informatics in the big data age: a survey," *ACM Computing Surveys,* vol. 49, no. 1, pp. 1-36, 2016.

[12] J. G. Greener, S. M. Kandathil, L. Moffat and D. T. Jones, "A guide to machine learning for biologists," *Nature Reviews Molecular Cell Biology,* vol. 23, no. 1, pp. 40-55, 2022.

[13] M. Li, G. Nanda and R. Sundararajan, "Evaluating Different Machine Learning Models for Predicting the Likelihood of Breast Cancer," *Advanced Aspects of Engineering Research,* vol. 2, pp. 132-142, 2021.

[14] S. Alghunaim and H. H. Al-Baity, "On the scalability of machine-learning algorithms for breast cancer prediction in big data context," *IEEE Access,* vol. 7, pp. 91535-91546, 2019.

[15] V. Anusuya and V. Gomathi, "An efficient technique for disease prediction by using enhanced machine learning algorithms for categorical medical dataset," *Information Technology and Control ,* vol. 50, no. 1, pp. 102-122, 2021.

[16] S. Jayatilake, M. Chinthaka and G. U. Ganegoda, "Involvement of machine learning tools in healthcare decision making," *Journal of healthcare engineering,* 2021.

[17] I. Eroglu, V. Sevilimedu, A. Park, T. A. King and M. L. Pilewskie, "Accuracy of the breast cancer surveillance consortium model among women with LCIS," *Breast Cancer Research and Treatment,* vol. 194, no. 2, pp. 257-264, 2022.

[18] D. A. Aljawad, E. Alqahtani, A.-K. Ghaidaa, N. Qamhan, N. Alghamdi, S. Alrashed, J. Alhiyafi and S. O. Olatunji, "Breast cancer surgery survivability prediction using bayesian network and support vector machines," in *International Conference on Informatics, Health & Technology*, 2017.

[19] A. Witteveen, G. F. Nane, I. M. Vliegen, S. Siesling and M. J. IJzerman, "Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence," *Medical decision making,* vol. 38, no. 7, pp. 822-833, 2018.

[20] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics,* vol. 2, 2006.

[21] J.-C. Lévesque, C. Gagné and R. Sabourin, "Bayesian hyperparameter optimization for ensemble learning," *arXiv preprint arXiv:1605.06394,* 2016.

[22] A. Yala, C. Lehman, T. Schuster, T. Portnoi and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology ,* vol. 292, no. 1, pp. 60-66, 2019.

[23] Z. Guo, L. Xu and N. A. Asgharzadeholiaee, "A Homogeneous Ensemble Classifier for Breast Cancer Detection Using Parameters Tuning of MLP Neural Network," *Applied Artificial Intelligence,* vol. 36, no. 1, 2022.

[24] Y. Mate and N. Somai, "Hybrid feature selection and Bayesian optimization with machine learning for breast cancer prediction," in *7th International Conference on Advanced Computing and Communication Systems*, 2021.

[25] G. Chugh, S. Kumar and N. Singh, "Survey on machine learning and deep learning applications in breast cancer diagnosis," *Cognitive Computation,* pp. 1-20, 2021.

[26] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro and A. Liotta, "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning," *Computer Methods and Programs in Biomedicine,* vol. 223, 2022.

[27] World Health Organization, "WHO position paper on mammography screening," WHO position paper on mammography screening, 2014.

[28] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 71, no. 3, pp. 209-249, 2021.

[29] J. L. Bernal, S. Cummins and A. Gasparrini, "Interrupted time series regression for the evaluation of public health interventions: a tutorial," *International journal of epidemiology,* vol. 46, no. 1, pp. 348-355, 2017.

[30] D. Oyewola, D. Hakimi, K. Adeboye and M. D. Shehu, "Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis," *International Journal of Engineering Technologiesq,* vol. 2, no. 4, pp. 142-145, 2016.

[31] L. Westerdijk and S. Bhulai, "Predicting malignant tumor cells in breasts," *Master Business Analytics ,* 2018.

[32] S. T. Kakileti, G. Manjunath, A. Dekker and L. Wee, "Robust estimation of breast cancer incidence risk in presence of incomplete or inaccurate information," *Asian Pacific Journal of Cancer Prevention,* vol. 21, no. 8, 2020.

[33] C. Hou, X. Zhong, P. He, B. Xu, S. Diao, F. Yi, H. Zheng and J. Li, "Predicting breast cancer in Chinese women using machine learning techniques: algorithm development," *JMIR medical informatics,* vol. 8, no. 6, 2020.

[34] K. Rajendran, M. Jayabalan and V. Thiruchelvam, "Predicting breast cancer via supervised machine learning methods on class imbalanced data," *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 8, 2020.

[35] F. S. Firouzabadi, A. Vard, M. Sehhati and M. Mohebian, "An optimized framework for cancer prediction using immunosignature," *Journal of medical signals and sensors ,* vol. 8, no. 3, 2018.

[36] National Cancer Institute-funded Breast Cancer Surveillance Consortium, "RISK ESTIMATION DATASET," 2006. [Online]. Available: https://www.bcsc-research.org/data/rfdataset.

[37] S. Sakri and S. Basheer, "Fusion Model for Classification Performance Optimization in a Highly Imbalance Breast Cancer Dataset," *Electronics ,* vol. 12, no. 5, 2023.

[38] Z. Peng, J. Wei, X. Lu, H. Zheng, X. Zhong and W. Gao, "Treatment and survival patterns of Chinese patients diagnosed with breast cancer between 2005 and 2009 in Southwest China: An observational, population-based cohort study," *Medicine ,* vol. 95, no. 25, 2016.

[39] C. Luo, X. Zhong, Y. Fan, Y. Wu, H. Zheng and T. Luo, "Clinical characteristics and survival outcome of patients with estrogen receptor low positive breast cancer," *The Breast ,* vol. 63, pp. 24-28, 2022.

[40] Y. Xie, L. Yang, Y. Wu, H. Zheng and Q. Gou, "Adjuvant endocrine therapy in patients with estrogen receptor-low positive breast cancer: A prospective cohort study," *The Breast,* vol. 66, pp. 89-96, 2022.

[41] J. E. Barrett, C. Herzog, A. Jones, O. C. Leavy, I. Evans, S. Knapp and D. Reisel, "The WID-BC-index identifies women with primary poor prognostic breast cancer based on DNA methylation in cervical samples," *Nature Communications,* vol. 13, no. 1, 2022.

[42] G. Alfian, M. Syafrudin, I. Fahrurrozi, F. T. D. A. Norma Latif Fitriyani, T. Widodo, N. Bahiyah, F. Benes and J. Rhee, "Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method," *Computers ,* vol. 11, no. 9, 2022.

[43] Y. Shieh, D. Hu, L. Ma, S. Huntsman, C. C. Gard, J. W. Leung and J. A. Tice, "Breast cancer risk prediction using a clinical risk model and polygenic risk score," *Breast cancer research and treatment,* vol. 159, pp. 513-525, 2016.

[44] M. Li and R. Sundararajan, "Application of Machine Learning Algorithms on Breast Cancer Dataset," in *Proc. 2018 Electrostatics Joint Conference*, 2018.

[45] C. Ming, V. Viassolo, N. Probst-Hensch, I. D. Dinov, P. O. Chappuis and M. C. Katapodi., "Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations," *British journal of cancer,* vol. 123, no. 5, 2020.

[46] D. M. Lang, J. C. Peeken, S. E. Combs, J. J. Wilkens and S. Bartzsch, "Deep learning based HPV status prediction for oropharyngeal cancer patients," *Cancers ,* vol. 13, no. 4, 2021.

[47] N. S. M. Ashokkumar, P. Anandan, M. Yaswanth, K. S. K. Bhanu Murthy, T. A. Alahmadi, S. A. Alharbi, S. S. Raghavan and S. A. Jayadhas, "Deep Learning Mechanism for Predicting the Axillary Lymph Node Metastasis in Patients with Primary Breast Cancer," *BioMed Research International,* 2022.

[48] H. Saleh, H. Alyami and W. Alosaimi, "Predicting breast cancer based on optimized deep learning approach," *Computational Intelligence and Neuroscience,* 2022.

[49] K. Uddin, M. Mohi, N. Biswas, S. T. Rikta and S. K. Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique," *Computer Methods and Programs in Biomedicine Update,* vol. 3, 2023.

[50] M. Botlagunta, M. D. Botlagunta, M. B. Myneni, D. Lakshmi, A. Nayyar, J. S. Gullapalli and M. A. Shah, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Scientific Reports,* vol. 13, no. 1, 2023.

[51] S. A. Kumar and A. K. Thangavelu, "Factors affecting the outcome of Global Software Development projects: An empirical study," in *International Conference on Computer Communication and Informatics*, 2013.

[52] C. Yang, J. Yang, Y. Liu and X. Geng, "Cancer Risk Analysis Based on Improved Probabilistic Neural Network," *Frontiers in computational neuroscience,* vol. 14, no. 58, 2020.

[53] B. Kurian and V. L. Jyothi, "Breast cancer prediction using an optimal machine learning technique for next generation sequences," *Concurrent Engineering,* vol. 29, no. 1, pp. 49-57, 2021.

[54] M. Savić, V. Kurbalija, M. Ilić, M. Ivanović, D. Jakovetić, A. Valachis, S. Autexier, J. Rust and T. Kosmidis, "Analysis of machine learning models predicting quality of life for cancer patients," in *Proceedings of the 13th International Conference on Management of Digital EcoSystems*, 2021.

[55] Z. Guan, T. Huang, A. M. McCarthy, K. Hughes, A. Semine, H. Uno, L. Trippa, G. Parmigiani and D. Braun, "Combining breast cancer risk prediction models," *Cancers ,* vol. 15, no. 4, 2023.

[56] F. Hamedani-KarAzmoudehFar, R. Tavakkoli-Moghaddam, A. R. Tajally and S. S. Aria, "Breast cancer classification by a new approach to assessing deep neural network-based uncertainty quantification methods," *Biomedical Signal Processing and Control ,* vol. 79, 2023.

[57] Leventi-Peetz, Anastasia-Maria and K. Weber, "Probabilistic machine learning for breast cancer classification," *Mathematical Biosciences and Engineering ,* vol. 20, no. 1, pp. 624-655, 2023.

[58] L. Hussain, S. Ansari, M. Shabir, S. A. Qureshi, A. Aldweesh, A. Omar, Z. Iqbal and S. A. C. Bukhari, "Deep convolutional neural networks accurately predict breast cancer using mammograms," *Waves in Random and Complex Media,* pp. 1-24, 2023.

[59] S. Kayikci and T. M. Khoshgoftaar, "Breast cancer prediction using gated attentive multimodal deep learning," *Journal of Big Data,* vol. 10, no. 1, pp. 1-11, 2023.

[60] C. Apté and S. Weiss, "Data mining with decision trees and decision rules," *Future generation computer systems ,* vol. 13, no. 2, pp. 197-210, 1997.

[61] R. G. Mantovani, T. Horváth, R. Cerri, S. B. Junior, J. Vanschoren, A. Carlos, P. d. Leon and F. d. Carvalho, "An empirical study on hyperparameter tuning of decision trees," *arXiv preprint arXiv:1812.02207 ,* 2018.

[62] J. D. Kelleher, B. M. Namee and A. D'Arcy, "Fundamentals of machine learning for predictive data analytics: algorithms," *Worked examples, and case studies,* 2015.

[63] Ç. Demirel, A. A. Tokuç and A. T. Tekin, "Click prediction boosting via Bayesian hyperparameter optimization based ensemble learning pipelines," *Intelligent Systems with Applications,* 2023.

[64] C. Zhang and Y. Ma, Ensemble machine learning: methods and applications, Springer Science & Business Media, 2012.

[65] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature biotechnology,* vol. 26, no. 9, pp. 1011-1013, 2008.

[66] D. Che, Q. Liu, K. Rasheed and X. Tao, "Decision tree and ensemble learning algorithms with their applications in bioinformatics," *Software tools and algorithms for biological systems,* pp. 191-199, 2011.

[67] Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita and H. Fujiyoshi, "Boosted random forest," *IEICE TRANSACTIONS on Information and Systems ,* vol. 98, no. 9, pp. 1630-1636, 2015.

[68] Y. Xia, C. Liu, Y. Li and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert systems with applications,* vol. 78, pp. 225-241, 2017.

[69] A. H. Aoulad, C. Mohamed, B. Abdelhamid, N. Ourdani and T. E. Alami, "A Comparative Evaluation use Bagging and Boosting Ensemble Classifiers," in *International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2022.

[70] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE transactions on neural networks and learning systems,* vol. 28, no. 10, pp. 2222-2232, 2016.

[71] M. Gao, G. Shi and S. Li, "Online prediction of ship behavior with automatic identification system sensor data using bidirectional long short-term memory recurrent neural network," *Sensors,* vol. 18, no. 2, 2018.

[72] Y. Benjamini, "Discovering the false discovery rate," *Journal of the Royal Statistical Society: series B (statistical methodology) ,* vol. 72, no. 4, pp. 405-416, 2010.

[73] A. J. Larner and A. J. Larner, "Paired Measures," *The 2x2 Matrix: Contingency, Confusion and the Metrics of Binary Classification ,* pp. 15-47, 2021.

[74] S. Das, A. Rai, M. L. Merchant, M. C. Cave and S. N. Rai, "A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies," *Genes,* vol. 12, no. 12, 2021.

[75] J. Muschelli III, "ROC and AUC with a binary predictor: a potentially misleading metric," *Journal of classification,* vol. 37, no. 3, pp. 696-708, 2020.

[76] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behavioural processes ,* vol. 148, pp. 56-62, 2018.