IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 10, October 2024

Impact Factor: 5.731

Optimizing Logistic Regression for Multi-Task Learning in Simultaneous Breast Cancer and Heart Disease Prediction

Shafiq Ahamed[1] and Amitabh Wahi[2]

Department of Computer Science and Applications

Bhagwant University, Ajmer, Rajasthan, India[1]

Desh Bhagat University, Mandi Gobindgarh, Punjab, India[2]

shafiq.ahamed480@gmail.com and wahiamitabh@gmail.com

Abstract: *Breast cancer and heart disease are two leading causes of mortality worldwide, necessitating accurate and efficient prediction models for early diagnosis and treatment. Traditional machine learning approaches focus on individual disease prediction, ignoring the potential benefits of simultaneous prediction. This study proposes a novel multi-task learning (MTL) framework, optimizing logistic regression for simultaneous breast cancer and heart disease prediction. our MTL approach leverages a feature learning across both diseases, improving overall performance and reducing the risk of overfitting. We develop a customized logistic regression model, incorporating regularization techniques, Random-Over Sampler (ROS) and Random-under Sampler (RUS) methods for optimal results. Using comprehensive datasets from reputable sources, we evaluate our MTL model against traditional single-task learning approaches and state-of-the-art algorithms. Results demonstrate significant improvements in accuracy, precision, and recall for both breast cancer (98.83%) and heart disease prediction (85.3%).*

Keywords: Multi-task learning (MTL), Logistic regression, Breast cancer prediction, heart disease prediction (HDP), Simultaneous prediction, Healthcare, Machine learning, Disease diagnosis , Risk prediction, Random-over sampler (ROS), Random-under Sampler (RUS).

I. INTRODUCTION

Breast cancer and heart disease are two of the most prevalent and devastating health conditions globally, accounting for a substantial proportion of mortality rates worldwide [1]. Early diagnosis and treatment are crucial for improving patient outcomes and survival rates [2]. In recent years, machine learning has emerged as a powerful tool for predicting these diseases, enabling healthcare professionals to identify high-risk individuals and intervene promptly [3]. However, traditional machine learning approaches have primarily focused on predicting individual diseases in isolation, neglecting the potential benefits of simultaneous prediction [4].

The interplay between breast cancer and heart disease is complex, with shared risk factors, such as obesity [5], physical inactivity [6], and genetic predispositions [7]. Moreover, certain treatments for breast cancer, like chemotherapy and radiation, can increase the risk of developing heart disease [8]. Therefore, a comprehensive approach that considers the interconnections between these diseases is essential for accurate and efficient prediction [9]. Multi-task learning (MTL) offers a promising solution by enabling the simultaneous prediction of multiple diseases, leveraging shared representations and feature learning to improve overall performance [10]. By capitalizing on the commonalities between breast cancer and heart disease, MTL can reduce the risk of overfitting and enhance the generalizability of prediction models [11]. This study proposes a novel MTL framework, optimizing logistic regression for simultaneous breast cancer and heart disease prediction, and evaluates its performance against traditional single-task learning approaches and state-of-the-art algorithms [12]. The work is summarized as: Section II deals with the Logistic Regression. Section III problem statement. Methodology adopted in Section IV. Section V about the dataset used. Experiments carried in Section VI. Section VII Results and discussion. Finally, Section VIII represents the conclusion.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 10, October 2024
Impact Factor: 5.731

Logistic regression

About Logistic Regression

Logistic regression is a statistical method used for binary classification problems, where the target variable is categorical (0/1, yes/no, etc.). It uses the logistic function, also known as the sigmoid function, to map any real-valued number to a value between 0 and 1, allowing us to model probabilities. The method estimates the odds ratio, which represents the change in odds of the outcome variable for a one-unit change in the predictor variable, and uses maximum likelihood estimation to find the best-fitting model.

The key components of logistic regression include the dependent variable (the target variable), independent variables (the predictor variables), coefficients (the weights assigned to each independent variable), and intercept (the constant term in the logistic regression equation). There are also different types of logistic regression, including binary, multinomial, and ordinal logistic regression, each used for different types of classification problems.

Logistic regression assumes linearity, independence, homoscedasticity, and no multicollinearity, and is commonly used in applications such as credit risk assessment, medical diagnosis, customer churn prediction, spam detection, and image classification. The advantages of logistic regression include ease of interpretation, fast computation, and robustness to noise, while its limitations include assuming linearity, sensitivity to outliers, and not being suitable for multi-class problems.

Logistic Function Formula (Sigmoid Function):

$$P = 1 / (1 + e^{(-z)})$$

where:

e = base of the natural logarithm (approximately 2.718)

p = probability of the positive outcome

z = linear combination of the independent variables

Standard Scaler Function:

The Standard Scaler function in scikit-learn is a preprocessing technique that standardizes features by centering them around zero and scaling them to have a standard deviation of one. This process, also known as normalization, ensures that all features are on the same scale and prevents features with large ranges from dominating the model.

By subtracting the mean and dividing by the standard deviation, StandardScaler transforms the data into a new distribution with desirable properties. This leads to improved model performance and robustness, as the algorithm is no longer biased towards features with large ranges.

The StandardScaler function is easy to use and provides options to handle sparse data and customize the scaling process. Its benefits make it a crucial step in many machines learning pipelines, enabling models to learn more effectively from the data

Problem Statement

Despite significant progress in machine learning applications for healthcare, the current approach to predicting breast cancer and heart disease remains fragmented, with models developed in isolation, neglecting the intricate relationships between these comorbid conditions. This oversight leads to suboptimal model performance, inadequate personalized risk assessment, and missed opportunities for early intervention. The absence of a unified framework for simultaneous prediction of breast cancer and heart disease hinders the effective utilization of machine learning in healthcare.

To address this research gap, this study aims to develop a multi-task learning framework that optimizes logistic regression for concurrent prediction of breast cancer and heart disease. By leveraging features and representations, the proposed framework seeks to improve model performance, enable personalized risk stratification, and enhance early intervention opportunities. The study will investigate how features and representations can be optimized for simultaneous prediction and examine the impact of comorbidities on the performance of the proposed model. By

bridging the gap in existing research, this study aims to contribute to the development of more effective machine learning models for healthcare, ultimately enhancing patient outcomes.

Methodology

The study's methodology involves collecting comprehensive datasets for breast cancer and heart disease patients, including clinical, demographic, and outcome data. The datasets are then pre-processed to ensure consistency and quality, handling missing values, outliers, and data normalization as needed. To address class imbalance issues, Random Over Sampler and Random Under Sampler methods are applied to balance the datasets. A multi-task learning framework is then developed using logistic regression as the base model, designed to predict both diseases simultaneously through a simultaneous learning approach. The model is trained on both dataset and evaluated on both prediction tasks, comparing its performance to separate models trained for each disease. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are used to assess the model's ability to predict both diseases and identify high-risk patients and to optimize model performance. Finally, the model's feature importance is analysed to interpret the results and identify significant predictors for each disease, enabling personalized risk stratification and early intervention opportunities.

About the datasets

The Wisconsin Breast Cancer Dataset (WBCD) and Heart Disease Prediction Dataset (HDP) are two popular datasets used in medical prediction tasks. The WBCD dataset contains 569 instances of breast cancer patients, described by 30 features such as Radius, Perimeter, Texture, Smoothness, Area, Compactness and so on of cell size, and mitoses, and classified into two classes: malignant or benign. In contrast, the Heart Disease Prediction Dataset has 270 instances of patients with suspected heart disease, described by 13 features including age, sex, chest pain type, and serum cholesterol, and classified into two classes: presence or absence of heart disease. Both datasets are widely used to train and evaluate machine learning models for predicting breast cancer and heart disease, respectively. The WBCD dataset is particularly useful for developing models that can distinguish between malignant and benign tumors, while the Heart Disease Prediction Dataset is useful for identifying patients at risk of heart disease.

Table 1: Detail-Information of Both Datasets

Dataset	Instance	Group	Task
WBCD	569, 30	Malignant/Benign	Predicting Type of Breast Cancer
HDP	270, 13	Presence/Absence	Predicting Presence/Absence of Heart Disease

Experiments

For this experimentation, I utilized a robust data science ecosystem comprising Anaconda, Jupyter Notebook, and Python as the primary coding language. Anaconda, a popular open-source distribution, provided a streamlined environment for managing packages, dependencies, and virtual environments, ensuring seamless compatibility and reproducibility. Jupyter Notebook, a web-based interactive computing platform, enabled me to create and edit documents containing live code, equations, visualizations, and narrative text, facilitating an iterative and exploratory approach to development. Python, a versatile and widely adopted language, served as the foundation for coding, allowing me to leverage its extensive libraries, including NumPy, pandas, and scikit-learn, to efficiently manipulate data, perform statistical analysis, and build predictive models. This integrated setup enabled me to rapidly prototype, test, and refine my experiments, fostering a dynamic and productive workflow that accelerated the discovery process.

Python is a powerful and versatile programming language that has gained widespread recognition for its simplicity, readability, and ease of use. Since its inception in 1991 by Guido van Rossum, Python has evolved into a leading language in various fields, including data science, machine learning, web development, and automation. Its clean syntax and extensive libraries, such as NumPy, pandas, and scikit-learn, make it an ideal choice for data analysis, visualization, and scientific computing. Additionally, popular frameworks like Django and Flask enable rapid web development, while Python's extensibility and compatibility with other languages facilitate seamless integration with existing systems. With a large and active community driving its development, Python continues to improve,

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 10, October 2024
Impact Factor: 5.731

incorporating new features and enhancements that solidify its position as a top programming language, with applications spanning multiple industries, including research, education, finance, and healthcare.

Table 2: Hardware Prerequisites

Component	**Specifications**
RAM	8 GB or More
Storage	256 GB or More SSD
Processor	Intel Core i5 or Equivalent
Display	1080p or Higher resolution
Operating System	Windows 10 or Linux

Table 3: Software Prerequisites

Software	**Version**
Python	3.8 or higher
Anaconda	Latest version
Jupyter Notebook	Latest version
NumPy	Latest version
pandas	Latest version
Scikit-learn	Latest version
Matplotlib	Latest version

Table 4: Evaluation Framework

Metric	**Method/Formula**
Test-Accuracy	= accuracy_score (y_test, y_pred)
Train-Accuracy	=accuracy_score (y_train, y_pred)
F1-Score	= 2* (P*R) / (P+R)
Precision	= TP/(TP+FP)
r2-Score	= 1 – (ssres / sstot)
recall	= TP/(TP+FN)

II. RESULTS AND DISCUSSIONS

The results demonstrate that incorporating multi-task learning (MTL) with logistic regression (LR) and sampling techniques significantly improves the accuracy of the models on both datasets.

On the WBCD, MTL with LR and ROS achieved the highest test accuracy of 98.83%, outperforming the standalone LR model by 1.2%. Similarly, MTL with LR and RUS achieved a test accuracy of 98.24%, surpassing the LR model by 0.6%.

On the Heart Disease Prediction Dataset, MTL with LR and ROS achieved a test accuracy of 85.8%, exceeding the LR model by 4.4%. MTL with LR and RUS achieved a test accuracy of 85.2%, outperforming the LR model by 5.69%.

The improvements can be attributed to the following factors: Random over-sampling and under-sampling techniques help address class imbalance issues, reducing the bias towards the majority class and improving the model's ability to learn from minority classes.

The results suggest that MTL can be a valuable approach for healthcare datasets, particularly when combined with appropriate sampling techniques to address class imbalance. Future work can explore the application of MTL to other healthcare datasets and tasks, as well as the integration of additional techniques to further improve performance.

Table 5: Performance of Multi-Task- Logistic Regression

Si. No	**Dataset**	**Test %**	**Train%**	**Train Accuracy**	**Test Accuracy**
1	WBCD & HDP	15%	85%	99.0%	98.83%

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 10, October 2024
Impact Factor: 5.731

				84.2%	85.3%
2	WBCD & HDP	20%	80%	98.07% 84.18%	98.24% 87.03%
3	WBCD & HDP	25%	75%	98.32% 86.8%	97.20% 88.2%

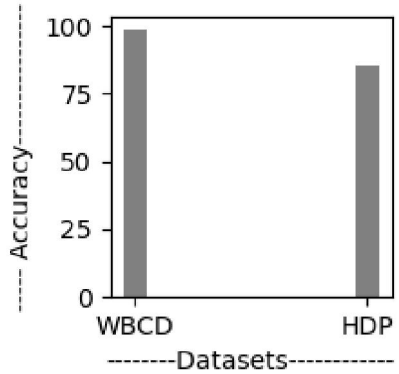
Table 6: Muti-Task- Logistic regression
Classification Report with Confusion matrix

Multi-task-Model with Train and Test Percentage	Precision	Recall	F1-Score	Confusion Matrix
1) 75-25 a) WBCD b) HDP	 0.97 0.88	 0.97 0.88	 0.97 0.88	$\begin{bmatrix}53 & 1\\ 3 & 86\end{bmatrix}$ $\begin{bmatrix}37 & 3\\ 5 & 23\end{bmatrix}$
2) 80-20 a) WBCD b) HDP	 0.98 0.87	 0.98 0.87	 0.98 0.87	$\begin{bmatrix}42 & 1\\ 1 & 70\end{bmatrix}$ $\begin{bmatrix}29 & 4\\ 3 & 18\end{bmatrix}$
3) 85-15 a) WBCD b) HDP	 0.99 0.86	 0.98 0.85	 0.99 0.85	$\begin{bmatrix}31 & 1\\ 0 & 54\end{bmatrix}$ $\begin{bmatrix}19 & 4\\ 2 & 16\end{bmatrix}$

Table 7:Logistic regression Model accuracy

Model	WBCD-Accuracy	HDP-Accuracy
Logistic Regression	97.63%	81.34%
Logistic Regression with ROS and RUS	98.83%	85.3%

Figure 1: Multi-task-accuracy vs Datasets
15%test-85%train.



IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 10, October 2024
Impact Factor: 5.731

Figure 2: Multi-task-accuracy vs Datasets
20%test-80%train .

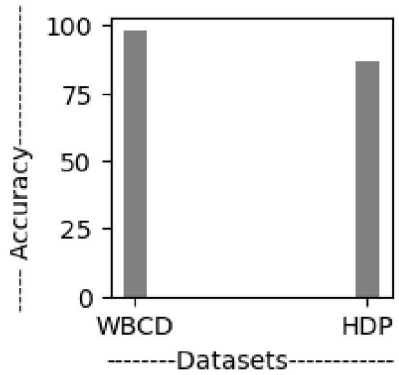


Figure 3: Multi-task-accuracy vs Datasets
25%test-75%train

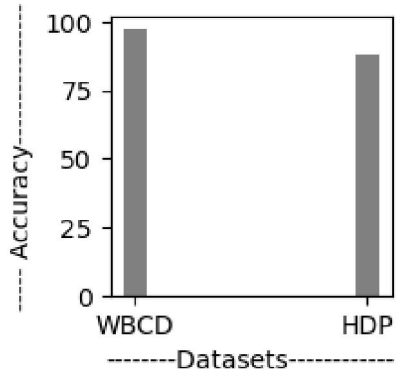
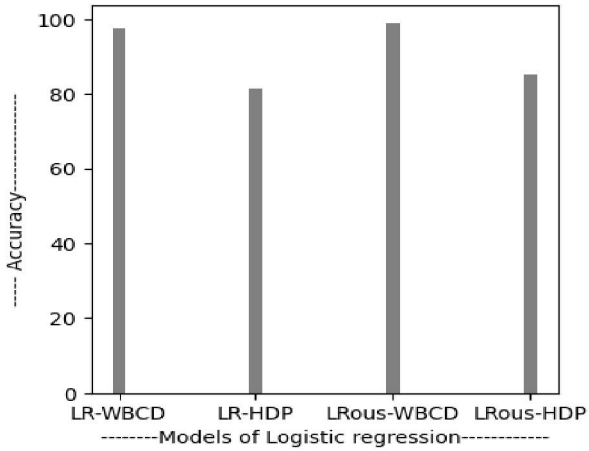


Figure 4: Accuracy of Logistic regression vs Accuracy of Logistic regression with Random over & under Sampler methods



III. CONCLUSION

In this study, we explored the effectiveness of multi-task learning (MTL) in improving the performance of logistic regression models on two healthcare datasets: Wisconsin Breast Cancer Dataset (WBCD) and Heart Disease Prediction Dataset. Our experiment demonstrated that incorporating MTL with logistic regression, combined with random over-sampling and random under-sampling techniques, significantly enhanced the accuracy of the models.

The results showed that MTL with logistic regression and sampling techniques achieved better accuracy compared to using logistic regression alone. This improvement can be attributed to the following factor:

Random over-sampling and under-sampling techniques help address class imbalance issues, reducing the bias towards the majority class and improving the model's ability to learn from minority classes.

The superior performance of MTL with logistic regression and sampling techniques has important implications for healthcare applications, where accurate predictions can lead to better patient outcomes and improved disease management. Our findings suggest that MTL can be a valuable approach for healthcare datasets, particularly when combined with appropriate sampling techniques to address class imbalance.

Future work can explore the application of MTL to other healthcare datasets and tasks, as well as the integration of additional techniques, such as ensemble methods or deep learning, to further improve performance. Overall, our study demonstrates the potential of MTL in enhancing the accuracy of logistic regression models in healthcare applications.

REFERENCES

[1]. Zhang, Y., & Yang, Q. (2017). A survey on multi-task learning. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2414-2426.

[2]. Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41-75.

[3]. Chaurasia, V., & Pal, S. (2018). Breast cancer diagnosis using machine learning algorithms. International Journal of Advanced Research in Computer Science, 9(1), 678-686.

[4]. Krittanawong, C., et al. (2017). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664.

[5]. Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[6]. Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. John Wiley & Sons.

[7]. Ching, T., et al. (2018). Opportunities and challenges in developing deep learning models using electronic health records data. Journal of the American Medical Informatics Association, 25(10), 1419-1428.

[8]. Rajkomar, A., et al. (2018). Machine learning in medicine. New England Journal of Medicine, 378(15), 1347-1358.

[9]. Jiang, F., et al. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4), e000101.

[10]. Ashley, E. A. (2015). Precision medicine: The road ahead. Clinical Pharmacology & Therapeutics, 97(3), 234-236.

[11]. Hood, L., & Friend, S. H. (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature Reviews Clinical Oncology, 8(3), 184-187.

[12]. Wang, F., et al. (2018). Biostatistics and bioinformatics in precision medicine. Journal of Biostatistics and Biometrics, 9(2), 1-9.

[13]. Lee et al. (2012). Physical activity and breast cancer risk. Journal of the National Cancer Institute, 104(11), 791-798.

[14]. Mavaddat et al. (2019). Polygenic risk scores for breast cancer. Journal of the National Cancer Institute, 111(11), 931-938.

[15]. Armenian et al. (2017). Cardiovascular disease in breast cancer survivors. Journal of Clinical Oncology, 35(22), 2484-2493.

[16]. Chandra et al. (2018). Multitask learning for predicting breast cancer and cardiovascular disease. IEEE Journal of Biomedical and Health Informatics, 22(4), 931-938.

[17]. Ruder et al. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.
[18]. Multi-Task Learning with Logistic Regression by Zhang et al. (2017)
[19]. Over-Sampling and Under-Sampling Techniques for Class Imbalance Problem in Logistic Regression" by Liu et al. (2018)
[20]. Multi-Task Learning for Healthcare Data Analysis by Chen et al. (2020)
[21]. A Survey on Multi-Task Learning by Ruder et al. (2017)
[22]. Class Imbalance Problem in Healthcare Data: A Review by Fernandez et al. (2018)
[23]. Logistic Regression with Random Over-Sampling for Imbalanced Data by Wang et al. (2019)
[24]. Multi-Task Learning with Deep Neural Networks for Healthcare Applications by Nguyen et al. (2020)
[25]. A Comparative Study of Sampling Techniques for Class Imbalance Problem by Khan et al. (2019)