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

Chapter 1

Introduction

Nowadays, data analysis is one of the most developing fields of computer science due to the fact that the size of datasets is exponentially increasing day after day. Cancer prediction is one of those fields, using data analysis and Machine Learning (ML) algorithms for the estimation of cancer [1][2][3]. ML techniques can improve the performance of cancer prediction, the estimation accuracy of which has increased significantly (15%-20%) due to using the ML algorithm during the last years [4]. Breast cancer prediction itself can be used to define those potentially high-risk women and guide them to improve their lifestyle, avoiding future therapy and costs [5].
Half of the cancer cases are caused by some known risk factors [6][7]. For breast cancer, many risk factors, such as early menarche, late menopause, obesity, age at first birth, and hormone therapy affect the exposure period of breast tissue to hormones that lead to cancer [8][9][10]. The main problem of cancer diagnosis and prediction is the huge amount of data that cannot be dealt with in the traditional manual method (physician's observations), and a more powerful speed approach is needed [11][12][13]. Fortunately, the rapid development in the computer science field has revealed the hidden information inside those datasets and provided health organizations with useful tools for diagnosing and predicting cancer [6][14][15][16]. Many previous breast cancer prediction tools were designed; using some known machine learning models (Support Vector machines (SVM), K-Nearest Neighbour (K-NN), Random Forests (RF), Decision Trees (DT), Neural Networks, Naïve Bayes and Logistic Regression (LR)) [17][18][19][20][21]. Some researchers used deep learning methodologies and fused them with image models, obtaining mammography breast image features, with the textual information of risk factors to improve the prediction model's accuracy [22]. Some of these researches got benefit of the parameter optimizations and ensemble learning methods to enhance the performance significantly [21][23][24].

1.1 Open-Ended Questions

Some previous studies used the well-known machine learning models, and few of them used the idea of ensemble learning or using a mix of many models. Many studies used the Breast Cancer Surveillance Consortium dataset (BCSC) dataset (partially or completely), but none of them analyzed the probabilistic distribution of this dataset. All previous studies introduced the cancer prediction problem using specific cancer prediction results (yes or no). In our study, a range-based cancer score will be computed based on 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 can the designed graphical user interface improve the quality of the breast cancer prediction tool?

1.2 Aims and Objectives

The main aim of this thesis is to build a range-based breast cancer prediction system based on machine learning algorithms. This thesis consists of five main parts, including five 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 third part aims to use the well-known LSTM deep learning architecture in the prediction of breast cancer in order to enhance performance. The fourth and fifth branch is performed on the original dataset but after applying the probabilistic model to get the target column in its range-based score.

1.3 Scope of work

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 thesis:

I have proposed a novel weighting methodology as the first state-of-art of breast cancer prediction. This method 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.

In the second branch, I introduced 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.

1.4 Workflow

Our proposed system can assist physicians, cancer researchers, and even individuals in predicting breast cancer at a very early stage. Additionally, it can serve as an early warning tool for the potential development of breast cancer. Unlike traditional models, our system provides a range-based probability of developing breast cancer, offering more nuanced insights than a simple yes or no warning.

Chapter 2

Research Methodology

2.1 Dataset

I used the BCSC dataset in this thesis. It includes 280660 records and 12 risk factors. Besides these risk factors, the dataset includes a variable called “count”, which holds the frequency of each record within the dataset, as mentioned in the BCSC dataset.

2.2 Proposed system

The thesis 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.

I proposed using the ensemble learning model in the next step in order to achieve the best performance. For the third part of this study, a fusion of the machine learning and deep learning methods is proposed. The outputs of the first two sections of this study are also proposed to be used as inputs or assistant methods for the third part of this study.

2.2.1 Ensemble machine learning and deep learning

Ensemble learning is a method in which many classifiers (models) are fused to build a huge powerful model. It has the advantage of using many classifiers to improve performance.

A fusion of ensemble learning and hyperparameters optimization has been given a lot of attention in the last few years [25]. The proposed ensemble method is the AdaBoost algorithm [26], while the learner type is the decision trees algorithm [27] [28].

At the first step of AdaBoost algorithm, the best promising feature is chosen as the root node, then the splitting process is applied based on a specific criterion.

Many learners are created and learned sequentially [29, 30].

So, in each step, a decision tree learner is chosen and fitted so that the error is forwarded to the next step and used to learn the next step learner [29].

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

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. 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.

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

In the fifth section, we continue to work with the same dataset as in the fourth section. The key distinction here is that we are focused on a regression task, which means that the target column retains its values without any merging of adjacent categories.
We introduce three regression models: Decision Trees Regression (DTR), Random Forest Regression (RFR), and K-NN Regression. Additionally, an ensemble combining these three models will be constructed.
To train these models, 80% of the dataset will be utilized, with the remaining 20% reserved for evaluation as a test set. We will assess performance by comparing the distribution of actual and predicted breast cancer scores.

2.5 Performance evaluation

Many evaluation metrics are computed to evaluate the trained ensemble model that has been trained using the sub dataset and the entire BCSC dataset.
Those metrics include the True Positive Rate (TPR), the False Negative Rate (FNR), the Positive Predictive Rate (PPR) and the False Discovery Rate (FDR) [31][32].

2.6 Used tools

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

Chapter 3

New Scientific Contribution and Thesis Points

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.

Related publications: J1

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.

Related publications: J2

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.

Related publications: C1

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.

Chapter 4

Application of the study

This thesis aim is to develop a new tool for predicting breast cancer based on BCSC risk factors. Further, the prediction tool can give the cancer risk as a percentage and not only a (0/1) value. I have taken into account the "count" variable for good estimation which indicates the number of times this case is repeated in medical domain. The another new option that have been done is the weighting mechanism, in which a weight number of each risk factor is assigned in order to enhance the performance. After getting the final optimizable decision tree classifier, the tool is built based on the trained model. The tool is designed using MATLAB App designer.

Since my proposed methodologies targeted the medical domain, there are numerous application possibilities of the proposed techniques:

Clinical Decision Support Systems. The developed models can be integrated into clinical decision support systems to help physicians in taking their decisions about breast cancer risk. By providing a quantitative analysis of the likelihood of cancer based on a range of scores, clinicians can make informed decisions regarding the need for further testing or intervention.

Specialized Medicine. Utilizing the precise predictions of my designed models, personalized treatment plans could be more effectively fit to individual patients. By considering the precise predictions of breast cancer risk scores, treatment plan can be better fit with the severity and specificity of each case.

Screening Program Optimization. My designed models could contribute to the refinement of breast cancer screening programs by defining patients with higher risk scores who may benefit from more frequent monitoring, which could lead to achieve an earlier detection and improved patient treatments.

Healthcare Resource Optimization. The predictive capabilities of my proposed models can guide healthcare administrators in effectively allocating resource. They could prioritize high-risk patients, and plan for necessary interventions and follow-up care.

Research Environment developing. The modified version of the BCSC dataset in my study provides a guide and start point for further research into cancer prediction and classification. They may use it to test new hypotheses, develop additional models, and contribute to the body of knowledge in breast cancer research.

Patient Risk Communication. The good visualization of risk scores and predictions can be utilized to communicate risks to potential patients in a more understandable method. This can help patients to understand the implications of their diagnostic results and the importance of next treatment steps.

Future Work

The main limitation of the current study is the dataset dependency. Our study is concentrated on one specific dataset (The BCSC dataset). However, future experiments can be performed on other datasets, taking into account more updatable risk factors since breast cancer data is updated every year. Future work can also focus on using the designed tool as a decision support tool to predict breast cancer, register statistical information and make some experiments to evaluate the retrieved data of the designed tool.

Acknowledgement

This research has received funding form PPCU supported by NKFIH, financed under Thematic Excellence Programme (TUDFO/51757-1/2019-ITM).
Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C), available at: http://www.bcsc-research.org/.

List of publications

[J1] Khozama S, Mayya AM. A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning. Information Technology and Control. 2022 Dec 12;51(4):757-70. Impact Factor **1.11**.
[J2] S. Khozama, A. Mayya, "Study the Effect of the Risk Factors in the Estimation of the Breast Cancer Risk Score Using Machine Learning", Asian Pacific Journal of Cancer Prevention, Vol. 22, no.11, pp.3543-3551, 2021. Impact Factor **1.892**.
Note: Asian pacific journal of cancer prevention and information technology and control are indexed in Scopus and Web of Science.

Conferences

[C1] Mayya, A. and Khozama, S., 2023, Breast Cancer Prediction Using Score-Level Fusion of Machine Learning and Deep Learning Models, Machine Learning Applications in Bioinformatics, January 2023 in Rome.
[C2] Mayya, A. and Khozama, S., 2020, December. A Novel Medical Support Deep Learning Fusion Model for the Diagnosis of COVID-19. In *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI)* (pp. 1-6). IEEE.
[C3] Khozama S, Nagy Z, Gáspári Z, “Accelerating Charged Single alpha-helix Detection on FPGA”. BIOTECHNO 2020, The Twelfth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, IARIA-2020, September 27, 2020.

References:

[1] Faith, M. F. (2020, June). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In Healthcare (Vol. 8, No. 2, p. 111). Multidisciplinary Digital Publishing Institute.
[2] Patil, S., Moafa, I. H., Mosa Alfaifi, M., et al. (2020). Reviewing the Role of Artificial Intelligence in Cancer. Asian Pacific Journal of Cancer Biology, 5(4), 189-99. https://doi.org/10.31557/apjcb.2020.5.4.189-199.
[3] Kamal, V., & Kumari, D. (2020). Use of Artificial Intelligence/Machine Learning in Cancer Research During the COVID-19 Pandemic. Asian Pacific Journal of Cancer Care, 5(S1), 251-3. https://doi.org/10.31557/apjcc.2020.5.S1.251-253.
[4] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.
[5] Colditz GA, Wei EK (2015). Risk prediction models: applications in cancer prevention. Curr. Epidemiol. Rep, **2**, 245-0.
[6] Ahmad AS, Mayya, AM (2020). A new tool to predict lung cancer based on risk factors. Heliyon, 6, e03402.
[7] LAKY Z (2020). Cancer prevention: Modifiable risk factors. Policy Department for Economic, Scientific and Quality of Life Policies IPOL. doi:10.2861/169496.
[8] American Cancer Society (2019). Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society.
[9] American Cancer Society. Breast Cancer Risk and Prevention (2019). Atlanta: American Cancer Society.
[10] American Cancer Society. Breast Cancer Fact Sheet (2020). Atlanta: American Cancer Society.
[11] Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C., Iyengar, S. S. Computational health informatics in the big data age: a survey. ACM Computing Surveys (CSUR), 2016, 49(1), 1-36.
[12] Greener, J. G., Kandathil, S. M., Moffat, L., Jones, D. T. A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology, 2022, 23(1), 40-55.
[13] Li, M., Nanda, G., Sundararajan, R., Evaluating Different Machine Learning Models for Predicting the Likelihood of Breast Cancer. Advanced Aspects of Engineering Research, 2021, 2, 132-142.
[14] Alghunaim, S., Al-Baity, H. H. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. IEEE Access, 2019, 7, 91535-91546.
[15] Anusuya, V., Gomathi, V. An efficient technique for disease prediction by using enhanced machine learning algorithms for categorical medical dataset. Information Technology and Control, 2021, 50(1), 102-122.
[16] Senerath, J., Don, M., Chinthaka, A., Ganegoda, G. U.: Involvement of machine learning tools in healthcare decision making. Journal of Healthcare Engineering, 2021, 1-20.
[17] Eroglu, I., Sevilimedu, V., Park, A., King, T. A., Pilewskie, M. L. Accuracy of the Breast Cancer Surveillance Consortium model among women with LCIS, 2021, 190(3), 1-20.
[18] Aljawad, D. A., Alqahtani, E., Ghaidaa, A. K., Qamhan, N., Alghamdi, N., Alrashed, S., Olatunji, S. O. Breast cancer surgery survivability prediction using bayesian network and support vector machines. In 2017 International Conference on Informatics, Health and Technology (ICIHT) 2017, 1-6.
[19] Annemieke, W., Nane, G. F., Vliegen, I. M., Siesling, S., IJzerman, M. J. Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence. Medical decision making, 2018, 38(7), 822-833.
[20] Cruz, J. A., Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2006, 2(1).
[21] Lévesque, J. C., Gagné, C., Sabourin, R. Bayesian hyperparameter optimization for ensemble learning. arXiv preprint arXiv:1605.06394, 2016.
[22] Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology, 2019, 292, 60-66.
[23] Guo, Z., Xu, L., Asgharzadeholiaee, N. A.: A Homogeneous Ensemble Classifier for Breast Cancer Detection Using Parameters Tuning of MLP Neural Network. Applied Artificial Intelligence, 2022, 36(2), 1-21.
[24] Mate, Y., Somai, N. Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, 1, 612-619).

References

[25] Kelleher JD, Namee B, D'arcy A. (2020). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.
[26] Zhang, C., Ma, Y.: Ensemble machine learning: methods and applications. Springer Science Business Media, 2012, https://doi.org/10.1007/978-1-4419-9326-7.
[27] Kingsford, C., Salzberg, S. L. What are decision trees. Nature biotechnology, 2008, 26(9), 1011-1013.
[28] Che, D., Liu, Q., Rasheed, K., Tao, X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. Software tools and algorithms for biological systems, 2011, 191-199.
[29] Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T., Fujiyoshi, H. Boosted random forest. EICE Transactions on Information and Systems, 2015, 98(9), 1630-1636.
[30] Xia, Y., Liu, C., Li, Y., Liu, N.: A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, 2017, 225-241.
[31] K. Greff, R. Srivastava K., J. Koutník, B. 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.
[32] Larner, A. J., Paired Measures. In The 2x2 Matrix, Springer, Cham, 2021, 15-47.
[33] Das, S., Rai, A., Merchant, M. L., Cave, M. C., Rai, S. N. A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. Genes, 2021, 12(12), 1-29.