UNIVERSITATEA BABEŞ-BOLYAI CLUJ-NAPOCA
FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ
SPECIALIZAREA: INFORMATICĂ ÎN LIMBA ROMÂNĂ

LUCRARE DE LICENŢĂ

LUCRARE DE LICENŢĂ

Dioșan Silvia Laura

Gradul

Baciu Ioana

2024

Contents

1 Introduction

Breast cancer is known to be the most common cancers among women [11]. The American Cancer Society's estimates for breast cancer in the United States alone for 2023 are that about 297,790 new cases of invasive breast cancer will be diagnosed in women and about 43,700 women will die from breast cancer [12]. Since 1989, breast cancer death rates have been decreasing, believed to be as a consequence of early detection and increased awareness, as well as better treatments. This progress seems to have slightly stopped [12].

The main objective of this project is to understand and implement ways to detect the presence of cancerous, benign, or precancerous tumours in the breast in an efficient way, using machine learning. This would minimize the time doctors spend studying thousands of breast screenings to label them accordingly and aid early detection.

2 The scientific problem addressed

3 Existing methods of detecting breast cancer

There are several articles related to studying various ways of implementing breast cancer detection using machine learning. There is “Machine Learning Techniques for Breast Cancer Prediction” by Varsha Nemade and Vishal Fegade from Mukesh Patel School of Technology Management and Engineering, NMIMS Shirpur Campus, India [1]. They have documented different ML classification techniques and evaluated each of them using different performance measure, such as accuracy, precision, and recall. These techniques include Näive Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbour and Decision Tree, the latter being found to have the highest accuracy, 97%.

The dataset used in their experiments was the WDBC dataset, which contain features from 569 digitized images of a fine needle aspirate of a breast mass [13]. The algorithm implemented uses features from the image, rather than the image itself, some of which include radius, texture, area, perimeter, smoothness, compactness, concavity, concavity points, symmetry, fractal dimension. The following charts represent the performance of the classification techniques used.

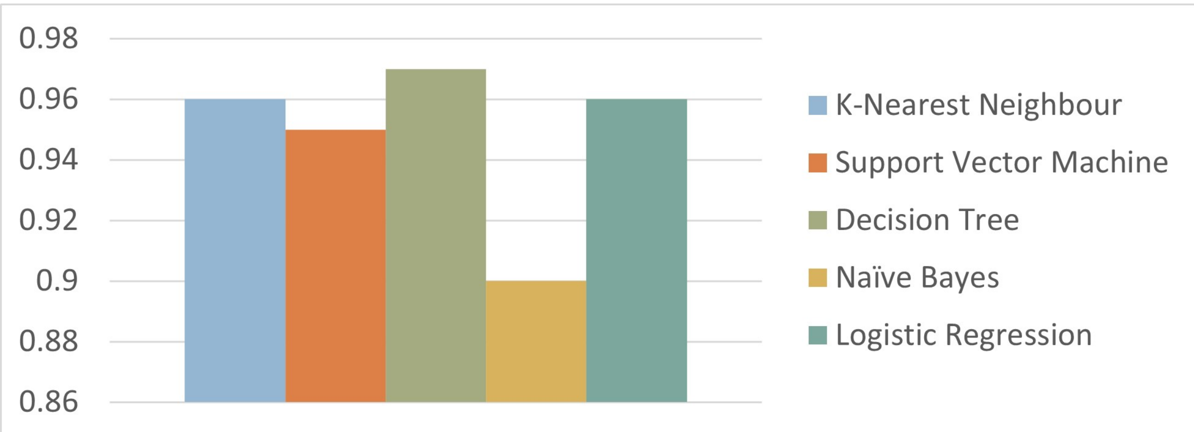

Figure 1: Accuracy of classification techniques

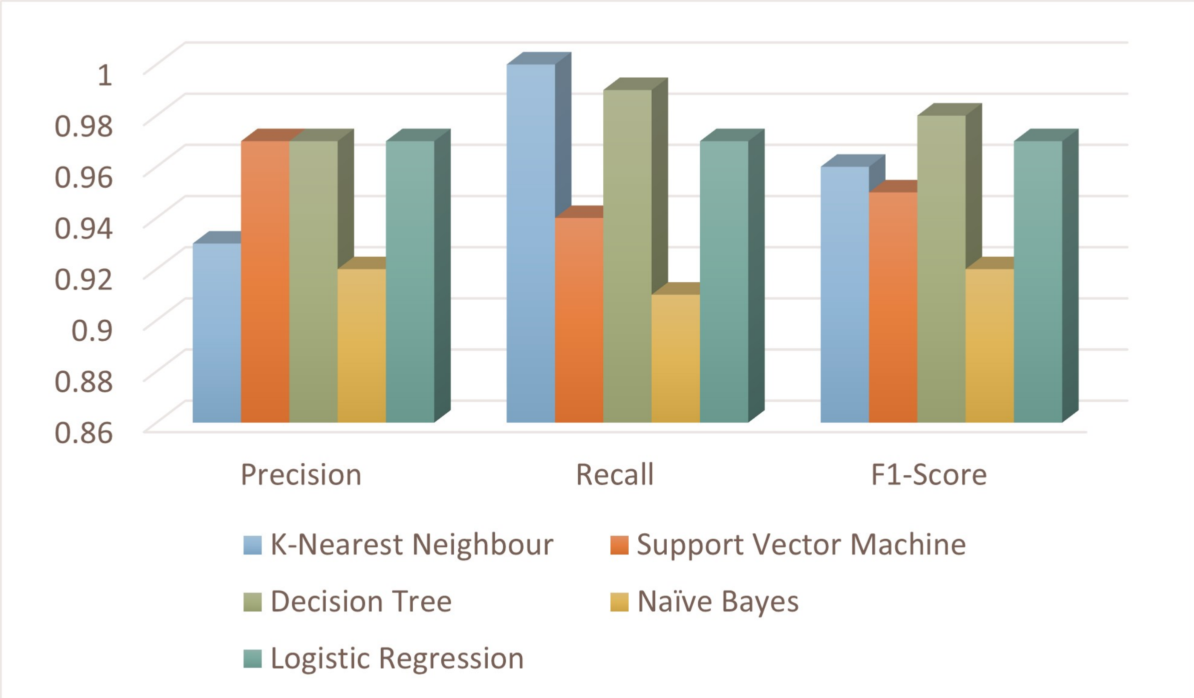

Figure 2: Performance of classification techniques for class benign

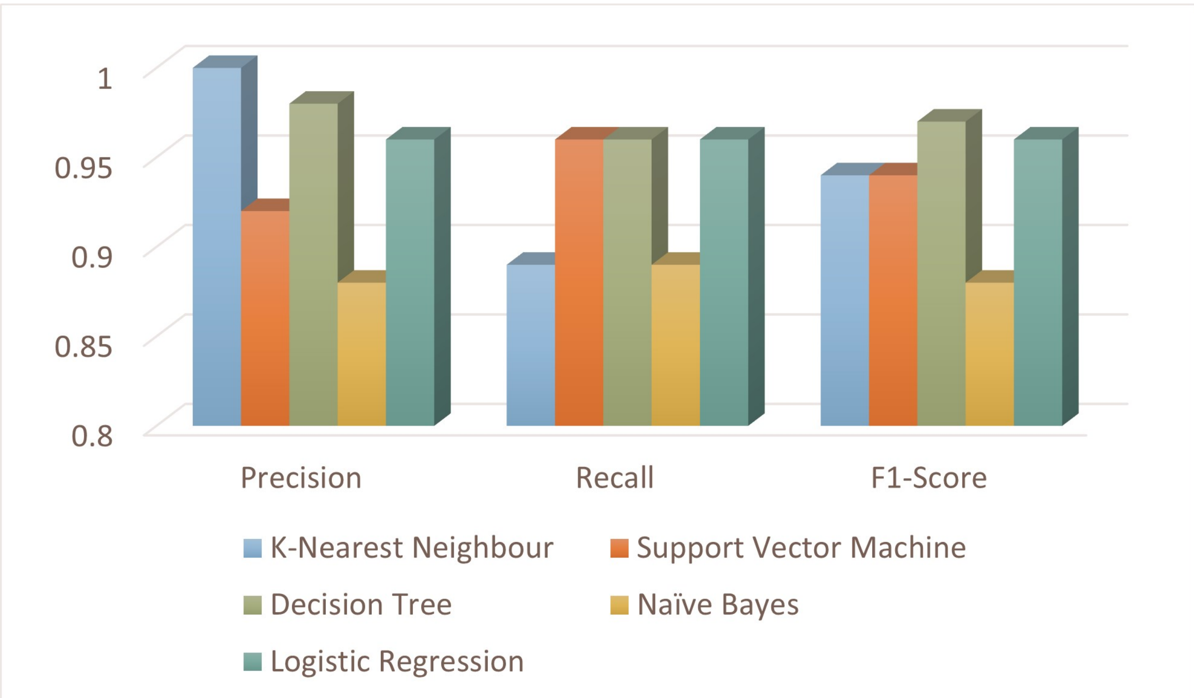

Figure 3: Performance of classification techniques for class malignant

Other relevant studies include “An enhanced Predictive heterogeneous ensemble model for breast cancer prediction" concluded by S. Nanglia et al. and got a 78% accuracy using KNN, SVM and DT [10]. Islam et al. got a 98.75% using ANN on the WDBC [5]. Amrane et al. proposed an approach using KNN and NB with 97.51% accuracy [3]. Dhahri et al. studied the usage of genetic programming techniques for selection of best features and parameter for the machine learning classifier [2].

4 Methods used in detecting breast cancer

4.1 Database used

The purpose of this thesis is to test the accuracy and precision resulting from training a GoogLeNet model on the Breast-Cancer-Screening-DBT database [14]. From this database, only a part of the data has been selecting for training and validating. Thus, the database I will be working on contains 719 screenings, of which 350 don't present any cancer, 280 have actionable skin neoplasms, 42 have benign tumors and 47 show signs of malignant cancerous tumors.

The Cancer Imaging Archive site provides a download link to a .tcia file that contains the screening files, which can be further downloaded using the NBIA Data Retriever. The images come in the form of .dcm files, which are DICOM files known in the medical forum, the abbreviation coming from Digital Imaging and Communications in Medicine. This format is different from other image formats because the information in grouped into data sets. For the sake of the patient's confidentiality, all sensitive information regarding the patient are removed before making the DICOM file available for the public. [8].

Also from the Cancer Imaging Archive website, files containing the file paths, ground truth labels and bounding boxes can be accessed. [4]. For reading the images from the .dcm files, a github repository has been provided by the publishers of the databse. [15]

4.2 GoogLeNet Architecture

4.3 EfficientDet Architecture

4.3.1 Main idea

The EfficientDet model is a new family of object detectors based on 2 main optimizations, those being efficient multi-scale feature fusion and model scaling. Along with those optimizations, different models have been tested to serve as the backbone of the model, and for this particular model, EfficientNet has proven to be a sufficient backbone architecture. [6]

The first challenge of this model was feature fusion, a concept introduced to aid convolutional neural networks in identifying objects from images, when the objects' sizes are either too small or exceed the convolutional kernel's receptive field. [9]. It has been discovered that changing the resolution of the image solved the sensitivity that neural networks have to the images' size, to some extend, as shown in figure 4

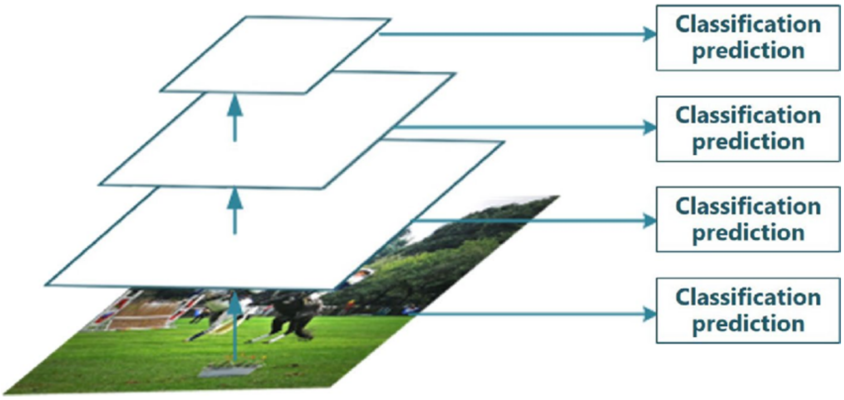

Figure 4: Pattern diagram of image pyramid

The problem with this discovery was that the cost in system storage of creating a pyramid of the same image, but in different resolutions and computational resources are to harsh, therefore deeming this method rarely usable. Because each resolution had it's advantages, a way to combine features from high-resolution features of shallow networks with the high-level semantic information of high-level network features was researched. Thus, the FPN, Feature Pyramid Network, has been used to extract from deep-layer networks and get the same features as the shallow-layer features by up-sampling, as shown in figure 5

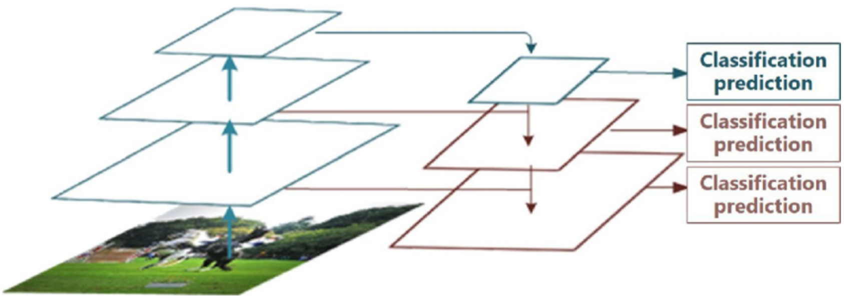

Figure 5: Pattern diagram of feature pyramid network

In this particular model, a BiPFN has been introduced, Bi-directional FPN, which utilizes learnable weights to learn the importance of different input features from different resolutions.

The second challenge is model-scaling. This method was popularized in [7], and is used to uniformly scale all dimensions of depth/width/resolution using a compound coefficient. This scaling method, called compound scaling, is looking at how the different dimensions interact with each other and scales them accordingly. For the EfficientDet architecture, this scaling method is used to scale up resolution, width and depth for all layers of the backbone, BiFPN and the bounding boxes and classification network.

Finally, for the backbone, EfficientNet has been observed to achieve better accuracy than other previously used backbones. [6]

4.3.2 Bi-directional Feature Pyramid Network

The baseline strategy for a standard FPN is to aggregate the features from the different levels of resolution in a top-down manner, from the features of the image with the lowest resolution to the ones from the highest one. For example, if a feature is extracted from the lowest resolution image, that feature is passed through a convolutional layer, used for feature processing, and the resulting output is then resized, which could be upsampling or downsampling in order the match the resolution of the following feature and so on. [6].

The conventional FPN is, however, limited by the one directional feature flow. To correct this use, PANet, Path Aggregation Network, is introduced and adds a path going bottom-up. This methods are still not ideal because of the cost of computations and parameters needed. The strategy proposed in [6] says that nodes that have only one edge going into it should be removed, because the lack of multiple input edges corresponds to less contribution to the overall feature network as it does not involve feature fusion. Moreover, between every input and output node at the same level, there are edges added in order to add more feature fusion without much cost, as it is on the same level. Finally, each bidirectional path is treated as one singular layer and more layers are involved in order to achieve more high-level feature fusion, as shown in figure 6

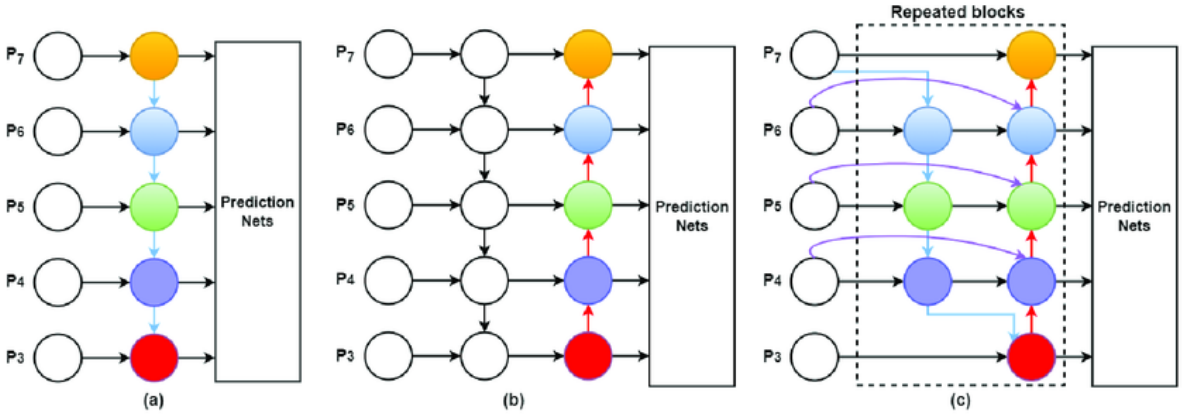

Figure 6: (a)-Conventional FPN; (b)-PANet; (c)-BiFPN

5 Experimental results obtained

6 Conclusions and possible improvements

7 Bibliography

References

[1] Naji M. A., El Filali S., Aarika K., Benlahmar E. H., and O. Abdelouhahid R. A.and Debauche. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191:487–492, 2021.

[2] Dhahri H., Al Maghayreh E., Mahmood A., Elkilani W., and Faisal Nagi M. *Automated breast cancer diagnosis based on machine learning algorithms.* Journal of healthcare engineering, 2019.

[3] Amrane M., Oukid S., Gagaoua I., and Ensari T. Breast cancer classification using machine learning. *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)*, pages 1–4, 2018.

[4] Buda M., Saha A., Walsh R., Ghate S., Li N., Świecicki A., Lo J. Y., and Mazurowski M. A. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA Netw Open*, 4, 2021.

[5] Islam M., Haque M., Iqbal H., Hasan M., Hasan M., and Kabir M. N. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1:1–14, 2020.

[6] Tan M., Pang R., and Le Q. V. Efficientdet: Scalable and efficient object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 0781–10790, 2020.

[7] Tan M. and Le Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 2020.

[8] Varma D. R. Managing dicom images: Tips and tricks for the radiologist. *Indian J Radiol Imaging*, 22:4–13, 2012.

[9] Lu S., Ding Y, Liu M., Yin Z., Yin L., and Zheng W. Multiscale feature extraction and fusion of image and text in vqa. *International Journal of Computational Intelligence Systems*, 16, 2023.

[10] Nanglia S., Ahmad M., Khan F. A., and Jhanjhi N. Z. An enhanced predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing and Control*, 72, 2022.

[11] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=64685580#6468558050a1e1bdf0de46de92128576e1d3e9b1.

[12] https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html.

[13] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

[14] https://www.cancerimagingarchive.net/collection/breast-cancer-screening-dbt/.

[15] https://github.com/mazurowski-lab/duke-dbt-data.