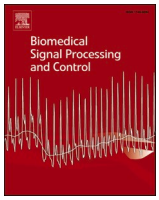
Biomedical Signal Processing and Control 72 (2022) 103279

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



An enhanced Predictive heterogeneous ensemble model for breast cancer prediction

S. Nanglia [a], Muneer Ahmad [a], Fawad Ali Khan [b], N.Z. Jhanjhi [c,*]

[a] Department of Information Systems, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[b] Department of Computer System & Technology, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[c] Scool of Computer Science and Engineering, SCE Taylor's University, Subang Jaya, Malaysia

ARTICLE INFO

Keywords:
Breast cancer
Machine learning
Data mining
Heterogeneous ensemble learning
Homogenous ensemble learning
Meta classifiers

ABSTRACT

Breast Cancer is one of the most prevalent tumors after lung cancer and is common in both women and men. This disease is mostly asymptomatic in the early stages thus detection is difficult, and it becomes complicated and expensive to be treated in later stages resulting in increased fatality rates. There are comparatively very few pieces of literature that investigated breast cancer employing an ensemble learning for cancer prediction as compared to single classifier approaches. This paper presents a heterogeneous ensemble machine learning approach, to detect breast cancer in the early stages. The proposed approach follows the CRISP-DM process and uses Stacking for building the ensemble model using three different algorithms – K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT). The performance of this meta classifier is compared with the individual performances of its base classifiers (KNN, SVM, DT) and other single classifiers – Logistic Regression (LR), Artificial Neural Network (ANN), Naïve Bayes (NB), Stochastic Gradient Descent (SGD) and a homogenous ensemble model of Random Forest (RF). The top 5 features – Glucose, Resistin, HOMA, Insulin, and BMI are derived by using Chi-Square. Evaluation of the model helps in estimating its consideration for early breast cancer prediction just by using the anthropometric data of humans. Performances of models are compared using metrics such as accuracy, AUC, ROC Curve, f1-score, precision, recall, log loss, and specificity using K-fold cross-validation of 2, 3, 5, 10, and 20 folds. The proposed ensemble model achieved the greatest accuracy of 78 % with the lowest log-loss of 0.56, at K = 20, thus rejecting the Null hypothesis. The derived p-value is 0.014, from the one-tailed *t*-test, which provides lower significance at $\alpha = 0.05$.

1. Introduction

Health is an important aspect of human life; thus, it is one of the major domains of computer science research and discoveries. From the many concerns in health care, the most difficult is the detection and treatment of cancer such as cervical cancer, blood cancer, breast cancer, neuroendocrine cancer, and so on. Recently, the most researched and feared topic in medical science is SARS–CoV–2 which has created havoc in our lives, but cancer is a pre-existing medical condition that is terminal and its existence along with the Covid-19 virus in a person's body could prove to be highly lethal. According to a recent research study, the infection rate of Covid-19, in cancer patients of France was 2.1% while the normal infection rate is 0.25% from March to April 2020[9]. Many countries advocate the early detection and treatment of tumors. According to[41], it has been measured that out of the estimated 19.3 million new cancer cases of the year 2020, around 2.3 million are cases of breast cancer. Quick detection and treatment of breast cancer are of utmost importance for saving lives and avoiding hefty treatment expenses. Early diagnosis and screening are the two strategies defined by the World Health Organization (WHO) for promoting uncovering of breast cancer in symptomatic and asymptomatic cases [14]. It could also be genetic,[44]reports that more than 5 to 6 percent of breast cancer patients are the results of genetic passing.

Tests for breast cancer include the screening of breasts to know the benign and malignant nature of the tumor. Mammography helps with a 30% reduction in mortality rate caused by breast cancer with sensitivity from 70% to 90% that depends on the quality of the image and experience of the radiologist[30]. Another approach for providing early detection is an analysis of the anthropometric data which could be collected from routine blood tests. There is a plethora of Machine and

* Corresponding author.
E-mail address: drnzamanj@gmail.com (N.Z. Jhanjhi).

https://doi.org/10.1016/j.bspc.2021.103279
Received 19 July 2021; Received in revised form 14 September 2021; Accepted 15 October 2021
Available online 4 November 2021