Project Design Phase-II

Open Source Frameworks

Date	06 May 2023
Team ID	NM2023TMID18418
Project Name	CancerVision: Advanced Breast Cancer Prediction with Deep Learning

An Ensembled Framework for Human Breast Cancer Survivability Prediction Using Deep Learning

Abstract:

Breast cancer is categorized as an aggressive disease, and it is one of the leading causes of death. Accurate survival predictions for both long-term and short-term survivors, when delivered on time, can help physicians make effective treatment decisions for their patients. Therefore, there is a dire need to design an efficient and rapid computational model for breast cancer prognosis. In this study, we propose an ensemble model for breast cancer survivability prediction (EBCSP) that utilizes multi-modal data and stacks the output of multiple neural networks. Specifically, we design a convolutional neural network (CNN) for clinical modalities, a deep neural network (DNN) for copy number variations (CNV), and a long short-term memory (LSTM) architecture for gene expression modalities to effectively handle multi-dimensional data. The independent models' results are then used for binary classification (long term > 5 years and short term < 5 years) based on survivability using the random forest method. The EBCSP model's successful application outperforms models that utilize a single data modality for prediction and existing benchmarks.

Introduction :

The human body is made up of approximately 30 trillion cells. Cancer originates from abnormal cell growth, resulting in the formation of a primary tumor [1]. Breast cancer predominantly affects women due to the excessive growth of breast cells. It is a highly invasive tumor and a leading cause of female fatalities [2]. In Pakistan, breast cancer is prevalent, with one out of every nine women at risk of the disease, and it has the highest cancer-related mortality rate [3]. According to the World Health Organization’s 2020 report, breast cancer is a significant cause of accidental death in women, with 8.2% of the Pakistani population dying from cancer. Figure 1 shows that breast cancer is the most commonly diagnosed cancer type, with a 28.7% diagnosis rate in 2020 [4]. Breast cancer has two types, including malignant and benign. Benign (non-invasive) cancer is a type of cancer that does not affect other organs. On the other hand, malignant (invasive) cancer spreads to neighboring tissues, making invasive cancer prognosis challenging due to varying clinical outcomes [5]. Thus, early and precise diagnosis and prognosis are crucial for timely decision making by physicians to improve patients’ survivability. Survivability can be categorized as short-term (<5 years) or long-term (>5 years). Prognostications aid physicians who work with short-term survivable patients with a multi-featured disease [6].

During the past few decades, the rapid growth of machine learning and deep learning

techniques with high throughput has provided deep insights into micro-arrays, gene expression,and clinical data

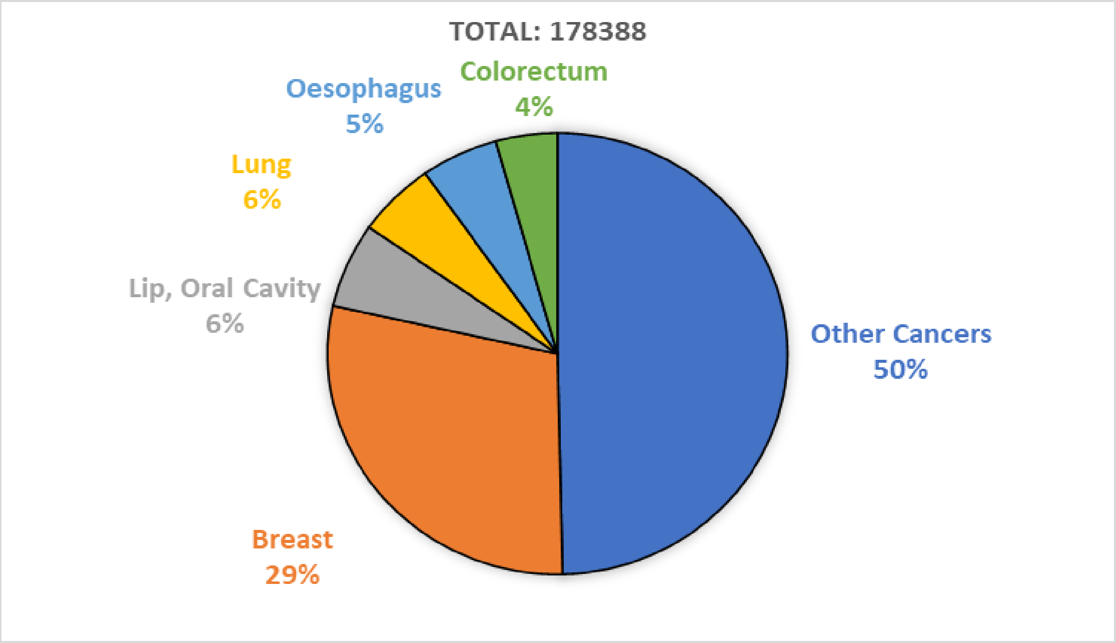

Figure 1. Breast cancer diagnosed in 2023; report by WHO.

This works relies on developing heterogeneous models based on a prognostic model

with stacking. The primary concern is to ensure the heterogeneity of the models for multimodal

data concerning the nature of the data. Different from prior works, we aim to

design an LSTM module for the feature extraction of gene expression data. The proposed

framework operates through three stages, specifically feature extraction, stacking, and

Classification.

Related Work:

Breast cancer prognosis is a critical need, alongside the prognosis of various other cancers. In pursuit of this objective, an ensemble approach that incorporates multiple machine learning models was investigated in [19]. The authors considered five different classification models and applied gene expression analysis to these models to obtain informative gene data. The proposed model was validated in lung cancer, stomach adenocarcinoma, and invasive carcinoma samples. The results confirmed the effectiveness of the proposed ensemble model. Another ensemble model was proposed in [20]. Here, the authors considered three classifiers, including support vector machines, logistics regression, and stochastic gradient descent optimization, for breast tumor classification. The proposed complex voting mechanism provided better results in comparison to existing benchmarks.

In [21], a multi-modal ensemble classification approach was investigated for human breast cancer prognosis. The authors proposed a deep learning-based stacked ensemble model using three datasets: CNV, clinical, and gene expression. They proposed a novel two-phase framework where features were extracted using a convolutional neural network in the first phase and RF was implemented in the second phase for output prediction. The results validated the effectiveness of the proposed multi-modal classification. However, there is still a need for a model that can effectively predict breast cancer patient prognosis and survivability. Existing benchmark models only work for a limited number of gene signatures or use similar neural networks for multi-modal data. Therefore, this work aims to design a heterogeneous model for multi-modal data to test the effectiveness of the proposed model in terms of accuracy.

Methods and Materials :

Dataset :

The METABRIC dataset is used for this work, extracted from 1980 valid patient records.

The METABRIC dataset comprises multi-dimensional data forms such as gene expression, copy number variation, and clinical information for breast cancer [21]. The total number of samples (patient count = 1980) was categorized into two subdivisions: long-term survivors (>five years) and short-term survivors (<5 years). The total number of samples comprises

1489 patient records for long-term survivors and 491 for short-term survivors. The remaining 64 patients of the total samples are alive, but the records of 3.2% of the total sample had incomplete five-year follow-up records. In these cases, we cannot determine whether the patients were long-term survivors or if they died within five years. Therefore, we continue our study by labeling these records as long-term survivors in EBCSP frameworks. This assumption is based on the very high survival chances reported by METABRIC for 64 unpublished patients. The duration of survival for the recorded patients was 125.1 months, whereas the median diagnosis age was 61 years. Using the survival threshold of five years, long-term survivors are labeled as ‘0’ and short-term survivors are labeled as ‘1’ for the binary classification model. The gene-expression (gene exp) and copy number variation (CNV) data have missing values, which are imputed using the weighted nearest neighbor algorithm [22].

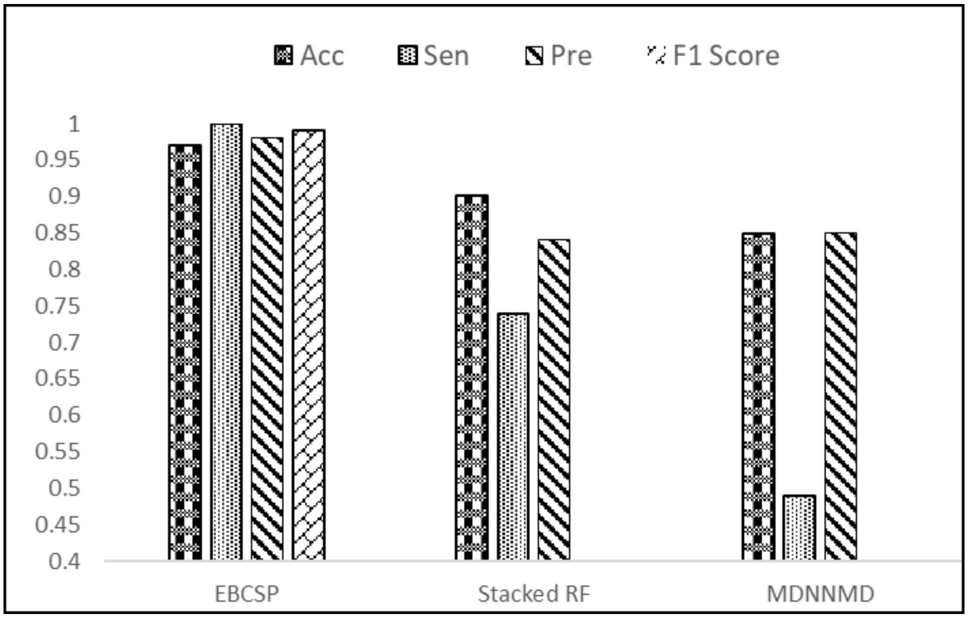

Figure 8. Result evaluation of EBCSP model with existing benchmarks