International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 529-533.
www.theijire.com ISSN No: 2582-8746

Cancer Diagnosis Using Machine Learning and Image Recognition

Muneesh Pal [1], Ahmed Ali Baig [2], Sahil Shikalgar [3], Shaikh Abdul Masood[4], Momin Sumaan[5]
[1]*Professor, Department of Information Technology Engineering, Armiet, Maharashtra, India.*
[2,3,4,5]*Department of Information Technology Engineering, Armiet, Maharashtra, India.*

How to cite this paper:
Muneesh Pal [1], Ahmed Ali Baig [2], Sahil Shikalgar [3], Shaikh Abdul Masood[4], Momin Sumaan[5], "Cancer Diagnosis Using Machine Learning and ImageRecognition", IJIRE-V4I02-529-533.

Abstract: *Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of MLtools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding ofcancer progression, an appropriate level of validation is needed in order for thesemethods to be considered in the everyday clinical practice. In this work, we present a review of recent ML approaches employed in the modeling of cancer progression. The predictive models discussed here are based on various supervised ML techniques as well as on different input features and data samples.Given the growing trend on the application of ML methods in cancer research, wepresent here the most recent publications that employ these techniques as an aimto model cancer risk or patient outcomes*

Key Word: *Cancer Diagnosis, Machine Learning, PyTorch Image Recognition, Artificial Neural Network, Inter Ductal Carcinoma.*

I.INTRODUCTION

Cancer research has seen a constant change throughout the last few decades. Scientists used several methods, suchas early-stage screening, to detect cancer kinds before they develop symptoms. Furthermore, they have created novel ways for predicting cancer therapy outcomes early on. Large volumes of cancer data have been collected and made available to the medical research community as a result of the introduction of new technologies in the field of medicine. However, reliable illness prediction is one of the most exciting and difficult jobs for clinicians. As a result, machine learning technologies have grown in popularity among medical researchers. These tools may discover and identify patterns and linksin complicated datasets, as well as accurately forecast future outcomes of a cancer type. Given the importance of personalized treatment and the growing trend of using ML techniques in cancer prediction and prognosis, we give a review of papers that use these methods. Prognostic and predictive factors are addressed in these investigations, which may be independent of a specific treatment or are incorporated to guide therapy for cancer patients.

Furthermore, we cover the sorts of ML approaches employed, the types of data they integrate, and the overall performance of each proposed scheme, as well as their advantages. An obvious trend in the proposed works includes the integration of mixed data, such as clinical and genomic. However, a common problem that we noticed in several works is the lack of external validation or testing regarding the predictive performance of their models. It is clear that the application of ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction.

Based on, the accuracy of cancer prediction outcome has significantly improved by 15% to 20% the last years, with the application of ML techniques. Artificial intelligence refers to computer programs, or algorithms, that use data to make decisions or predictions. To build an algorithm, scientists might create a set of rules, or instructions, for the computer to follow so it can analyze data and make a decision. With other artificial intelligence approaches, like machine learning, the algorithm teaches itself how to analyze and interpret data. As such, machine learning algorithms may pick up on patterns that are not readily discernable to the human eye or brain. And as these algorithms are exposed to more new data, their ability to learn and interpret the data improves. Several studies have been reported in the literature and are based on differentstrategies that could enable the early cancer diagnosis and prognosis Specifically, these studies describe approaches relatedto the profiling of circulating miRNAs that have been proven a promising class for cancer detection and identification.

However, these methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumors. Various aspects regarding the prediction of cancer outcome based on gene expression signatures are discussed above. These studies list the potential as well as the limitations of microarrays for the prediction of cancer outcome. Even though gene signatures could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can beused in clinical practice, studies with larger data samples and more adequate validation are needed. In the present work only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

II.METHODOLOGY

- **Architectural Diagram**

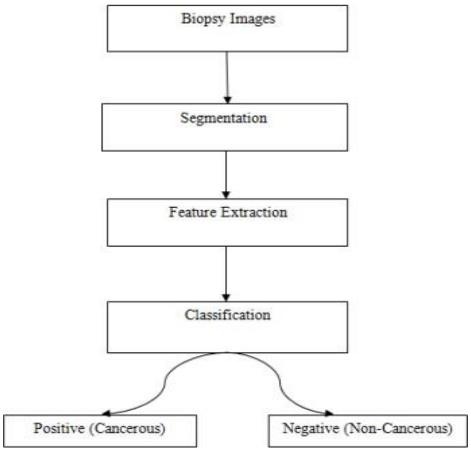


Fig:1 Architectural Diagram of cancer detection

Architectural diagram contains various steps:

- Microscopic tested image is taken as input after undergoing biopsy. The images are enhanced before segmentation to remove noise.
- Segmentation is done based on the input images which contains nuclei, cytoplasm and other features. They are segmented on the basis of region, threshold or a cluster and particular algorithms are applied.
- In feature extraction, various biologically interpretable and clinically notable shape and morphology based features are extracted from the segmented images which include grey level texture features, color based features, color grey level texturefeatures, Law's Texture Energy (LTE) based features, Tamura's features, and wavelet features.
- Finally the images are classified using Naive Bayes classifier.

Procedure methodology

AI is an umbrella term describing the mimicking of human intelligence by computers . Machine learning (ML), a subdivisionof AI, refers to training computer algorithms to make predictions based on experience, and can be broadly divided into supervised (where the computer is allowed to see the outcome data) or unsupervised (no outcome data are provided) learning. Both approaches look for data patterns to allow outcome predictions, such as the presence or absence of cancer, survival rates or risk groups. When analyzing unstructured clinical data, an often-utilised technique, both in oncology and more broadly, is natural language processing (NLP). NLP transforms unstructured free-text into a computer-analysable format, allowing the automation of resource-intensive tasks.

III.IMPLEMENTATION

- **Preparation of the gathered data set**

First things first, we have to install some libraries so that our program works. Here is a list of the libraries we will install: pandas, NumPy, Sklearn and seaborn. PyTorch as in torch has to be installed so that torch vision can work. After the installation is completed, let's import them into our code editor. Matplotlib is already included in Python that's why we can import it without installing it. All the below coding and outputs are expected.

Python Libraries

1. **NumPy** (short for Numerical Python) is a Python package for performing scientific computing with Python. It provides support for large, multi-dimensional arrays and matrices, as well as a large library of mathematical functions to operate on these arrays.

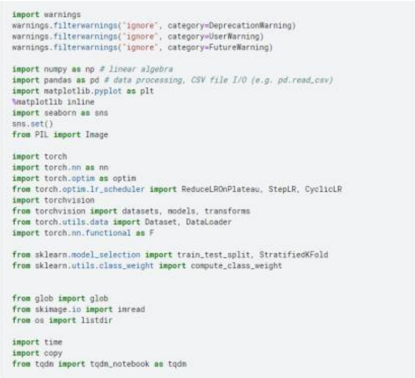
2. **Pandas** is a popular Python library for data manipulation and analysis. It provides data structures for efficiently storing and manipulating large datasets, as well as a variety of tools for working with data.

3. **Matplotlib** is a Python library used for data visualization. It provides a wide range of tools for creating high- quality plots, charts, and other visualizations.

4. **Seaborn** is a Python library for data visualization based on Matplotlib. It provides a high-level interface for creating

informative and attractive statistical graphics. Seaborn is built on top of Matplotlib and integrates well with Pandas data frames.

5. **Scikit-learn** (or sklearn) is a popular Python library for machine learning. It provides a range of tools for buildingpredictive models, including classification, regression, and clustering algorithms, as well as tools for data preprocessing, feature selection, and model evaluation.

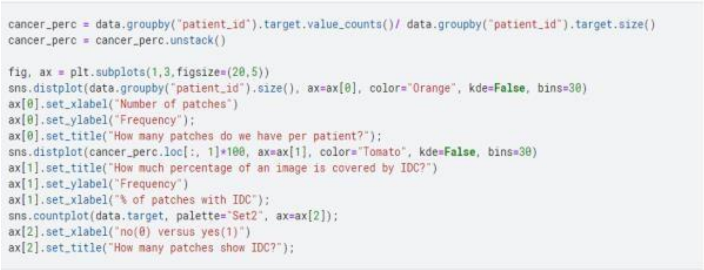


```
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
warnings.filterwarnings("ignore", category=UserWarning)
warnings.filterwarnings("ignore", category=FutureWarning)

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set()
from PIL import Image

import torch
import torch.nn as nn
import torch.optim as optim
from torch.optim.lr_scheduler import ReduceLROnPlateau, StepLR, CyclicLR
import torchvision
from torchvision import datasets, models, transforms
from torch.utils.data import Dataset, DataLoader
import torch.nn.functional as F

from sklearn.model_selection import train_test_split, StratifiedKFold
from sklearn.utils.class_weight import compute_class_weight


from glob import glob
from skimage.io import imread
from os import listdir

import time
import copy
from tqdm import tqdm_notebook as tqdm
```

Fig Importing Libaries

- **Exploratory Analysis**

We will start loading the data into a data frame, it is a good practice to take a look at it before we start manipulating it. This helps us to understand that we have the right data and to get some insights about it. The first thing we'll do to get some understanding of the data is using the head method. When you call the head method on the data frame, it displays the first five rows of the data frame. After running this method, we can also see that our data is sorted by the date index.



```
cancer_perc = data.groupby("patient_id").target.value_counts()/ data.groupby("patient_id").target.size()
cancer_perc = cancer_perc.unstack()

fig, ax = plt.subplots(1,3,figsize=(20,5))
sns.distplot(data.groupby("patient_id").size(), ax=ax[0], color="Orange", kde=False, bins=30)
ax[0].set_xlabel("Number of patches")
ax[0].set_ylabel("Frequency");
ax[0].set_title("How many patches do we have per patient?");
sns.distplot(cancer_perc.loc[:, 1]*100, ax=ax[1], color="Tomato", kde=False, bins=30)
ax[1].set_title("How much percentage of an image is covered by IDC?")
ax[1].set_ylabel("Frequency")
ax[1].set_xlabel("% of patches with IDC");
sns.countplot(data.target, palette="Set2", ax=ax[2]);
ax[2].set_xlabel("no(0) versus yes(1)")
ax[2].set_title("How many patches show IDC?");
```

Fig Exploratory Code

Insights

1. The number of image patches per patient varies a lot! This leads to the questions whether all images show the sameresolution of tissue cells of if this varies between patients.

2. Some patients have more than 80 % patches that show IDC! Consequently, the tissue is full of cancer or only a part of the breast was covered by the tissue slice that is focused on the IDC cancer. Does a tissue slice per patient cover the wholeregion of interest?

3. The classes of IDC versus no IDC are imbalanced. We have to check this again after setting up a validation strategy andfind a strategy to deal with class weights (if we like to apply them).

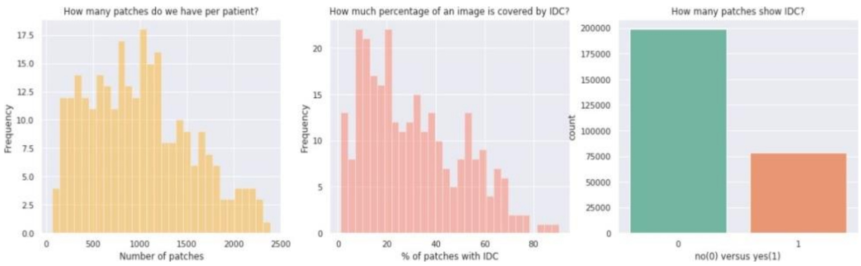
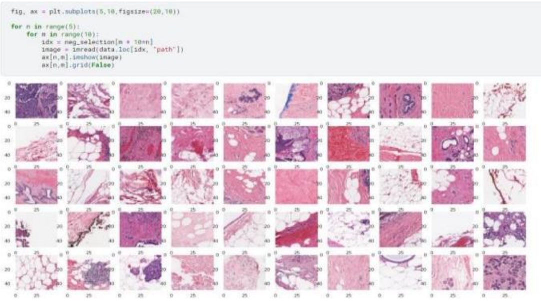


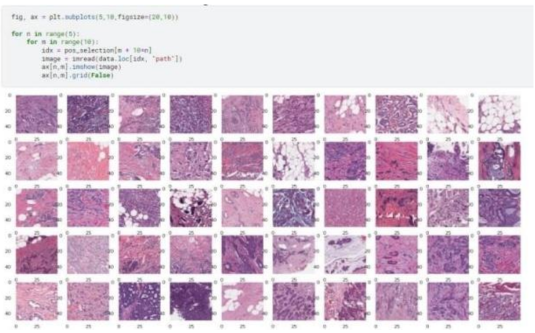
Fig output of the above code

Looking at healthy Patches and Cancer Patches: The function created below differentiates between the healthy patches and cancerous patches with the help of random forest classifier as positive and negative this is to train the model for further proceedings.



```
fig, ax = plt.subplots(5,10,figsize=(20,10))

for n in range(5):
    for m in range(10):
        idx = neg_selection[m + 10*n]
        image = imread(data.loc[idx, "path"])
        ax[n,m].imshow(image)
        ax[n,m].grid(False)
```

Fig a healthy patches



```
fig, ax = plt.subplots(5,10,figsize=(20,10))

for n in range(5):
    for m in range(10):
        idx = pos_selection[m + 10*n]
        image = imread(data.loc[idx, "path"])
        ax[n,m].imshow(image)
        ax[n,m].grid(False)
```

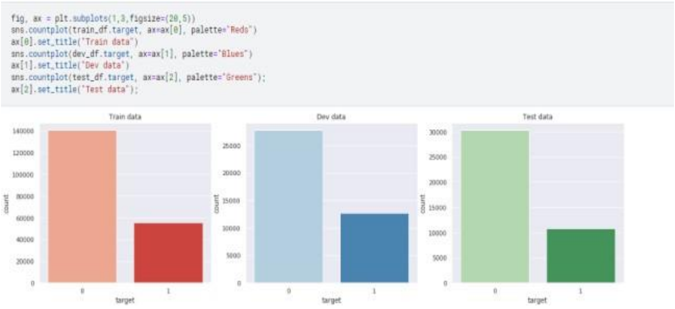
Fig Cancerous patches

Insights

- Sometimes we can find artifacts or incomplete patches that have smaller size than 50x50 pixels.
- Patches with cancer look more violet and crowded than healthy ones. Is this really typical for cancer or is it more typical for ductal cells and tissue?
- Though some of the healthy patches are very violet coloured too!
- Would be very interesting to hear what criteria are important for a pathologist.
- I assume that the holes in the tissue belong to the mammary ducts where the milk can flow through.

IV.RESULTS

let's start very simple by selecting 30 % of the patients as test data and the remaining 70 % for training and developing.This seems arbitrary and we should rethink this strategy in the next cycle of our data science workflow.



```
fig, ax = plt.subplots(1,3,figsize=(20,5))
sns.countplot(train_df.target, ax=ax[0], palette="Reds")
ax[0].set_title("Train data")
sns.countplot(dev_df.target, ax=ax[1], palette="Blues")
ax[1].set_title("Dev data")
sns.countplot(test_df.target, ax=ax[2], palette="Greens");
ax[2].set_title("Test data");
```

Fig Target distribution of train, test and development

We can see that the test data has more cancer patches compared to healthy tissue patches than train or dev.

We should keep this in mind!

In the present review, the most recent works relevant to cancer prediction/prognosis by means of ML techniques are presented. After a brief description of the ML branch and the concepts of the data pre-processing methods, the feature selection techniques and the classification algorithms being used, we outlined three specific case studies regarding the prediction of cancer susceptibility, cancer recurrence and cancer survival based on popular ML tools. Obviously, there is alarge amount of ML studies published in the last decade that provide accurate results concerning the specific predictive cancer outcomes. However, the identification of potential drawbacks including the experimental design, the collection of appropriate data samples and the validation of the classified results, is critical for the extraction of clinical decisions.

Moreover, it should be mentioned that in spite of the claims that these ML classification techniques can result in adequate and effective decision making, very few have actually penetrated the clinical practice. Recent advances in omics technologies paved the way to further improve our understanding of a variety of diseases; however more accurate validation results are needed before gene expression signatures can be useful in the clinics.

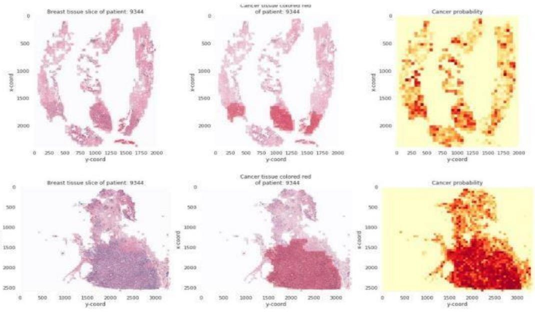

Breast tissue slice of patient: 9344
Cancer tissue colored red of patient: 9344
Cancer probability
Breast tissue slice of patient: 9344
Cancer tissue colored red of patient: 9344
Cancer probability
x-coord
y-coord

Fig Presumption of cancer detected by the model.

The above images classify weather the patient is not only having cancer but also the possibility of it spreading in over the span of time. After this we need to do some more work and generate the loss function to see weather the model is accurate to its core or not we will be doing that by evaluating the loss function and generating the accuracy of every single epoch and making a function to store it in a excel file so that it is easy for us to analyse the correctness of that model. So first lets start be setting up loss function and evaluation matrix also creating pytorch data loaders for the generation of the excel file A growing trend was noted in the studies published the last 2 years that applied semi-supervised ML techniques for modelling cancer survival. This type of algorithms employs labelled and unlabelled data for their predictions while it has been proven that they improved the estimated performance compared to existing supervised techniques. SSL can be though as a great alternative to the other two types of ML methods (i.e., supervised learning and unsupervised learning) that use, in general, only a few labelled samples.

V.CONCLUSION

"Prevention is better than cure". If successful, the model would help to detect cancer risk to individual and save a number of lives In this review, we discussed the concepts of ML while we outlined their application in cancer prediction/prognosis.

Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain. Red and unstructured big data sources. To achieve our mission towards precision oncology and better understand the complex mechanisms of cancer, intervention actions should be designed by means of evidence-based decision support tools to prevent what is preventable, optimize diagnostics and treatment and support the quality of life of patients and caregivers. Furthermore, considering the pandemic in the last two years and the situation in the public healthcare systems, we can admitthat cancer patients faced a severe and anxious period of follow-up visits trying to avoid a possible diagnosis which resulted in reduced hospitalizations and procedures.

As a result, we would foresee the influence of the pandemic on early cancer detection, on top of worsening prognosis and patient screening.

References

1. *D. Hanahan, R.A. Weinberg Hallmarks of cancer: the next generation Cell, 144 (2011), pp. 646-674*
2. *M.-Y.C. Polley, B. Freidlin, E.L. Korn, B.A. Conley, J.S. Abrams, L.M. McShane Statistical and practical considerations for clinical evaluation of predictive biomarkers J Natl Cancer Inst, 105 (2013), pp. 1677-1683*
3. *J.A. Cruz, D.S. Wishart Applications of machine learning in cancer prediction and prognosis Cancer Informant, 2(2006), p. 59*
4. *O. Fortunato, M. Boeri, C. Verri, D. Conte, M. Mensah, P. Suatoni, et al. Assessment of circulating microRNAs in plasma of lung cancer patients Molecules, 19 (2014), pp. 3038-3054*
5. *L. Ein-Dor, I. Kela, G. Getz, D. Givol, E. Domany Outcome signature genes in breast cancer: is there a unique set?Bioinformatics, 21 (2005), pp. 171-178*
6. *L. Ein-Dor, O. Zuk, E. Domany Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer Proc Natl Acad Sci, 103 (2006), pp. 5923-5928*
7. *T. Ayer, O. Alagoz, J. Chhatwal, J.W. Shavlik, C.E. Kahn, E.S. Burnside Breast cancer risk estimation with artificial neural networks revisited Cancer, 116 (2010), pp. 3310-3321*
8. *J.C. Platt, N. Cristianini, J. Shawe-Taylor Large margin DAGs for multiclass classification (1999), pp. 547-553*
9. *D. Cicchetti Neural networks and diagnosis in the clinical laboratory: state of the art Clin Chem, 38 (1992), pp. 9-10*
10. *A.J. Cochran Prediction of outcome for patients with cutaneous melanoma Pigment Cell Res, 10 (1997), pp. 162-167*