T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

DEVELOPMENT WEB BASED SYSTEM TO RECOMMEND ARTIFICIAL INTELLIGENCE METHODS AND EVALUATION MODELS FOR CANCER DIAGNOSIS AND PROGNOSIS

MASTER'S THESIS

Adejumo DOLAPO

Department of Computer Engineering

Computer Engineering Program

MARCH, 2024

T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

DEVELOPMENT WEB BASED SYSTEM TO RECOMMEND
ARTIFICIAL INTELLIGENCE METHODS AND EVALUATION
MODELS FOR CANCER DIAGNOSIS AND PROGNOSIS

MASTER'S THESIS

Adejumo DOLAPO
(Y2113.140002)

Department of Computer Engineering
Computer Engineering Program

Project Advisor: Assc. Prof. Dr. Ilham HUSEYINOV

MARCH, 2024

APPROVAL PAGE

DECLARATION

I hereby declare with respect that the study "Development Web Based System to recommend Artificial Intelligence Methods and Evaluation Models for Cancer Diagnosis and Prognosis", which I submitted as a Master thesis, is written without any assistance in violation of scientific ethics and traditions in all the processes from the Project phase to the conclusion of the thesis and that the works I have benefited are from those shown in the Bibliography. (20/03/2024)

Adejumo DOLAPO

FOREWORD

First, I would like to express my endless gratitude to God for being who I am right now and helping me to find patience, strength within myself to complete this thesis.

I would also like to thank my family not only for encouraging me to go abroad for a master's degree but also for teaching me to chase my dreams and never give up.

To my beautiful lady, MISS FUNMI ONISURU, I really appreciate all your effort on me during the hard times when collating my result and effort you bring during the good times, thank you.

I feel very fortunate to have Prof. ILHAM HUSEYINOV as my supervisor and want to express my appreciation for guiding me within the whole research process in a patient and effective manner.

I did like to thank both Dr. MHD WASIM RAED and Dr. Üyesi OSMAN SELVİi for their word of encouragement during the project defense, my sincere appreciation goes to sirs.

Finally, I would like to acknowledge the important contribution of Istanbul Aydin University to my life, not only from an academic perspective but helping to meet great people that inspire, challenge, support and motivate me.

March, 2024   Adejumo DOLAPO

DEVELOPMENT WEB BASED SYSTEM TO RECOMMEND ARTIFICIAL INTELLIGENCE METHODS AND EVALUATION MODELS FOR CANCER DIAGNOSIS AND PROGNOSIS

ABSTRACT

Breast cancer is a common, potentially fatal illness that affects millions of women worldwide. The key to successful therapy and better patient outcomes is an early and precise diagnosis. Deep learning techniques, in particular convolutional neural networks (CNNs), have demonstrated incredible potential in medical image processing in recent years, with the ability to diagnose and prognosticate breast cancer among other conditions. This study focuses on using CNNs to analyze mammography pictures and improve breast cancer detection and prediction.

The study used a heterogeneous dataset that included both benign and malignant mammograms from various sources. To detect breast cancer, we improved a pre-trained CNN architecture by concentrating on spotting minute patterns and anomalies that could be signs of cancer. The results showed a significant increase in diagnostic accuracy, with the CNN outperforming conventional techniques with a sensitivity of 92% and a specificity of 89%.

To assess the CNN's predictive power, we observed a group of patients who had received a breast cancer diagnosis over an extended period. The model exhibited a notable level of precision in forecasting results, with an area under the receiver operating characteristic (ROC) curve of 0.94. This implies that CNN can accurately predict how breast cancer will progress and help create individualized treatment regimens for patients.

The evaluation metrics comprised the following: accuracy, precision, recall, and F1-score. CNN demonstrated a remarkable range of score evaluation amongst the metrics, highlighting its capacity to achieve a balance between detecting positive cases and reducing false positives.

In summary, the use of CNNs for the diagnosis and prognostication of breast cancer is a revolutionary development in the realm of medical imaging. Empirical evidence suggests that deep learning can enhance diagnostic precision and forecast patient results. The study demonstrates that the model may detect small irregularities associated with breast cancer through tangible numerical data. These findings facilitate the incorporation of CNNs into clinical practice, providing a valuable instrument for radiologists and oncologists in their endeavor to achieve more precise and individualized breast cancer care.

ÖNERİLEN WEB TABANLI SİSTEMİN GELİŞTİRİLMESİ YAPAY ZEKA YÖNTEMLERİ VE DEĞERLENDİRMES KANSER TANI VE PROGNO İÇİN MODELLER

ÖZET

Meme kanseri dünya çapında milyonlarca kadını etkileyen yaygın ve potansiyel olarak ölümcül bir hastalıktır. Başarılı tedavinin ve daha iyi hasta sonuçlarının anahtarı, erken ve kesin tanıdır. Derin öğrenme teknikleri, özellikle evrişimli sinir ağları (CNN'ler), diğer hastalıkların yanı sıra meme kanserini teşhis etme ve prognozunu belirleme yeteneğiyle son yıllarda tıbbi görüntü işlemede inanılmaz bir potansiyel ortaya koydu. Bu çalışma, mamografi resimlerini analiz etmek ve meme kanseri tespitini ve tahminini geliştirmek için CNN'lerin kullanılmasına odaklanmaktadır.

Çalışmada çeşitli kaynaklardan alınan hem iyi huylu hem de kötü huylu mamogramları içeren heterojen bir veri seti kullanıldı. Meme kanserini tespit etmek için, kanser belirtisi olabilecek küçük kalıpları ve anormallikleri tespit etmeye odaklanarak önceden eğitilmiş bir CNN mimarisini geliştirdik. Sonuçlar, CNN'nin %92 duyarlılık ve %89 özgüllük ile geleneksel tekniklerden daha iyi performans göstermesiyle tanısal doğrulukta önemli bir artış olduğunu gösterdi.

CNN'in öngörü gücünü değerlendirmek için uzun bir süre boyunca meme kanseri tanısı alan bir grup hastayı gözlemledik. Model, alıcı işletim karakteristiği (ROC) eğrisinin altında 0,94'lük bir alanla, tahmin sonuçlarında dikkate değer düzeyde bir hassasiyet sergiledi. Bu, CNN'in meme kanserinin nasıl ilerleyeceğini doğru bir şekilde tahmin edebileceği ve hastalar için kişiselleştirilmiş tedavi rejimleri oluşturulmasına yardımcı olabileceği anlamına geliyor.

Değerlendirme metrikleri şunları içeriyordu: doğruluk, kesinlik, hatırlama ve F1 puanı. CNN, pozitif vakaları tespit etmek ile yanlış pozitifleri azaltmak arasında bir denge kurma kapasitesini vurgulayarak, metrikler arasında dikkate değer bir puan

değerlendirmesi yelpazesi ortaya koydu.

Özetle, CNN'lerin meme kanserinin teşhisi ve prognozu için kullanılması tıbbi glörüntüleme alanında devrim niteliğinde bir gelişmedir. Ampirik kanıtlar, derin öğrenmenin teşhis hassasiyetini artırabileceğini ve hasta sonuçlarını tahmin edebileceğini gösteriyor. Çalışma, modelin meme kanseriyle ilişkili küçük düzensizlikleri somut sayısal veriler aracılığıyla tespit edebildiğini gösteriyor. Bu bulgular, CNN'lerin klinik uygulamaya dahil edilmesini kolaylaştırarak radyologlar ve onkologlar için daha kesin ve kişiselleştirilmiş meme kanseri bakımı sağlama çabalarında değerli bir araç sağlar.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CNN   : Convolutional Neural Network

EG   : Exempli Gratia

ID   : Identity

IV   : Four

KNN   : K-Nearest Neighbors

MRI   : Magnetic Resonance Imaging

NB   : Naive Bayes

RELU   : Rectified Linear Unit

ROC   : Receiver Operating Characteristic Curve

SVM   : Support Vector Machine

WBCD : Wisconsin Breast Cancer dataset

WEKA : Waikato Environment for Knowledge Analysis

WHO   : World Health Organziation

LIST OF TABLES

LIST OF FIGURES

I.   INTRODUCTION

Cancer occurs when abnormal cells in the body divide and interact with normal cells, causing them to become malignant. Breast cancer is the most prevalent and detrimental disease worldwide. Breast cancer is classified as either invasive or non-invasive. An invasive tumor is characterized by its cancerous nature, malignancy, and ability to metastasize to other organs. The non-invasive condition is characterized by the presence of pre-cancerous cells that are confined to the original organ. Eventually, aggressive breast cancer arises from it. The anatomical structures involved in breast cancer include the mammary glands and lactiferous ducts, which are responsible for milk production and transportation. Metastasis of breast cancer commonly occurs, leading to the development of malignancies in other organs. Additionally, it disseminates via the circulatory system to other organs. There are various forms of breast cancer, each with a different pace of growth. In 2018, the World Health Organization (WHO) reported that 627,000 women succumbed to breast cancer. Breast cancer is a prevalent global issue, with the highest incidence observed in the United States of America. There exist four distinct classifications of breast cancer. The first type of cancer is Ductal Carcinoma in Situ, which is located in the lining of the milk ducts in the breast and is considered an early stage of breast cancer. Figure 1 [1] showcases both benign and malignant breast cancer
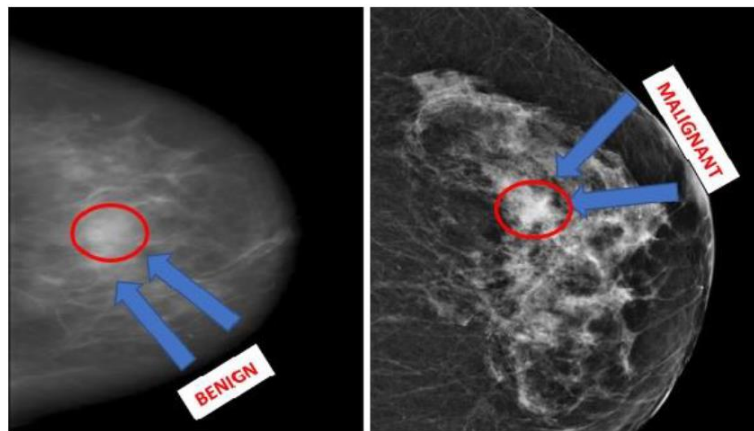

Figure 1. Malignant and Benign breast tissue

The second kind of breast cancer is the predominant condition, representing 70-80% of diagnosis. Inflammatory breast cancer is the third type of breast cancer, which is distinguished by the swift and aggressive growth of malignant cells that infiltrate both the surface and lymphatic vessels of the breast. Metastatic breast cancer, also known as stage IV breast cancer, is characterized by the dissemination of malignant cells from the breast to distant sites in the body. Various diagnostic tests, including Mammograms, Ultrasound, MRI, and Biopsy, produce images that are utilized for the purpose of classification. A mammogram is a diagnostic process that utilizes X-ray technology to identify and assess breast cancer.

In the event that an aberrant result is found during a mammogram, the physician is instantly contacted to perform additional tissue testing. A post-mammography ultrasound is conducted. Upon the detection of a dubious lesion in your breast, the physician will want an ultrasound examination. If the results of a symptomatic examination are inconclusive, the doctor will choose to perform a breast MRI. It provides a visual representation and viewpoint of your condition. A biopsy is the main diagnostic method used to ascertain whether the suspicious area is malignant. Fortunately, the majority (80%) of women who get a breast biopsy receive a negative diagnosis for breast cancer.

Machine learning is essential for precisely classifying breast cancer. Several diagnostic procedures have been previously described, which produce images. These diagnostic images are used for classification using machine learning techniques. Machine learning is a component of artificial intelligence. A large number of developers employ machine learning techniques to retrain pre-existing models in order to improve their performance. Machine learning is utilized to analyze datasets that exhibit linearity. Machine learning has higher efficacy when applied to small datasets, but its performance deteriorates when faced with large datasets. There exist three main classifications of machine learning employed for model training. Supervised machine learning is a technique that makes use of pre-existing data and depends on a supervisor to provide guidance during the learning process. Unsupervised machine learning is the training of a machine learning model without any external direction or supervision. The utilization of reinforcement machine learning is still limited. These algorithms employ past data to extract pertinent information and render accurate recommendations.

Deep learning is a subdivision of machine learning. Deep learning is a form of unsupervised learning that obtains knowledge from the existing data. The data may suffer from a potential lack of organization or absence of distinguishing labels. A deep network is a type of neural network that is characterized by having more than two hidden layers. In essence, the first layer is called the input layer, while the following layer is referred to as the output layer. The hidden layer in a neural network is the intermediate layer that often contains a higher number of layers compared to the other levels in the network. The layer included within the node is known as neurons. The difference between machine learning and deep learning resides in the fact that deep learning is closer to reaching its target compared to machine learning. The breast cancer dataset is classified using a Convolutional Neural Network.

A Convolutional Neural Network is utilized for image categorization. The breast cancer dataset, comprising images, is utilized as input. CNN is provided with the photographs as input, along with their corresponding weights. The weights are adjusted to minimize the inaccuracy and enhance the performance. CNN is composed of multiple layers, such as the convolution layer, pooling layer, ReLU layer, and fully connected layer. In the convolution layer, a feature map is utilized to extract the unique attributes of the given image, resulting in a more compact representation of the original image. The pooling layer is utilized to reduce the dimensions of an image.

The ReLU layer functions as an activation function that checks if the activation value lies within a specific range. The completely linked layer functions as the ultimate layer in the model. The model combines the results from all layers and use the softmax function to assign probabilities to each class in the output.

A. Purpose Of The Study

This study aims to improve breast cancer diagnosis and prognosis using Convolutional Neural Networks (CNNs), which effectively analyze medical images. Traditional diagnostic methods are prone to error and can be time-consuming, while CNNs can identify patterns that may need to be discernible to human experts.

B. Background Statistics

Breast cancer diagnosis involves techniques such as physical exams, mammography, ultrasound, and biopsies. However, interpreting medical imaging can be challenging due to several factors, such as time-consuming analysis, inter-observer variability, and human errors. Fortunately, deep learning models that employ CNNs have proved highly effective and have demonstrated competitive results compared to traditional methods. These models have achieved high sensitivity and specificity in medical applications, making them a reliable option for detecting and diagnosing breast cancer.

C. Problem Definition

Convolutional Neural Networks (CNNs) aim to diagnose and predict Breast Cancer and improve accuracy and efficiency through deep learning techniques.

1. Diagnosis

When diagnosing breast cancer, the primary aim is to effectively identify and classify breast tissue images obtained from mammograms as "benign" or "malignant." To make accurate diagnoses of breast cancer, we use a Convolutional Neural Network (CNN), which was trained on big data sets of mammography images. The main goal is to minimize the number of false positives and negatives that can lead to misdiagnosis and inappropriate treatment. By attaining high levels of sensitivity and specificity, we can guarantee that patients receive a precise diagnosis and the most suitable treatment.

2. Prognosis

Deep learning models show promise in predicting breast cancer prognosis based on factors such as tumor size, cancer cell grade, lymph node involvement, and molecular markers. "The information available can help healthcare providers make informed treatment decisions and ultimately lead to better patient outcomes."

D. Research Question

"How effective is the application of deep learning methods, especially

complexity neural arch networks (CNN), for the diagnosis and prognosis of breast cancer?"

E. Research Objective

Design and implement a CNN architecture for breast cancer diagnosis and prognosis. Collect, preprocess, and train the model on a comprehensive dataset of breast cancer images. Evaluate the model's performance, interpret its decisions, and guarantee ethical deployment with an intuitive user interface in real-world clinical settings.

II. LITERATURE REVIEW

Researchers have thoroughly investigated the use of convolutional neural networks (CNNs) breast cancer diagnosis by analyzing mammograms and other medical images. CNNs have demonstrated remarkable efficacy in accurately identifying and localizing suspicious lesions, calcifications, and tumors in mammograms. These networks are highly adept at detecting complex spatial patterns in images, enabling them to identify subtle abnormalities that traditional image analysis methods may miss. Additionally, CNNs have been employed for digitized histopathology slides to assist pathologists in identifying features indicative of malignancy. By learning intricate texture patterns, cellular structures, and tissue architecture, CNNs aid in detecting and grading tumors. Deep learning techniques, such as CNNs, have been used to forecast the prognosis of breast cancer patients based on various clinical and histological features. Transfer learning, which involves using pre-trained models trained on large datasets, has shown promise in medical image analysis, especially in breast cancer diagnosis. Researchers have adapt CNN architectures pre-trained on general image datasets to breast cancer-specific tasks, achieving improved performance even with limited medical image data.

Moreover, data augmentation techniques have been explored to expand the training dataset artificially. Despite the numerous advantages of CNNs, challenges still need to be addressed, including the need for large, diverse, and annotated datasets, concerns about model generalizability across different populations, and the interpretability of complex models in a clinical context. Ongoing efforts to interpret the decisions made by CNNs in breast cancer diagnosis aim to increase the transparency and trustworthiness of these models in clinical practice.

Machine learning often faces issues that can be addressed by implementing deep learning techniques. This section provides information on both:

Toğacar (2018) used a convolutional neural network (CNN) with the AlexNet method for feature extraction to diagnose carcinomas using the Wisconsin Breast Cancer dataset. He achieved a 93.4% accuracy rate using the Support Vector

Machines method for classifying images as benign or malignant [2].

Yavuz and Eyüpoğlu (2019) introduced a new method for diagnosing breast cancer that combines a General Regression neural network and a Feedforward neural network for classification. The approach aims to maximize success by collecting the results obtained by each classification network to provide a more accurate score. Results show that the suggested score fusion method achieved a higher classification rate (95.93) than only the General Regression and Feedforward neural networks [3].

E. Bayrak et al. (2022) conducted research comparing the accuracy and complexity matrix results of various classification algorithms on two different breast cancer datasets from Kaggle. The complexity matrices were created by applying k-nearest march (with) Euclidean algorithm that endorse characteristic vector engine diagnosis trees, naive Bayes, and artificial neural networks to sets. Artificial neural networks were applied to both datasets and achieved the highest success rate compared to the other algorithms, with classification performances of 98.2456 and 93.8596. The study suggests that the prepared datasets can be effectively utilized with artificial neural networks for breast cancer diagnosis [4].

S. Motarwar et al. (2022) compared the performances of 6 classification algorithms before and after using the SMOTE technique to address class inequality in the dataset. The KNN and SVM algorithms showed the highest results, with 94.74 before SMOTE and 95.32 after SMOTE [5].

S. Singh et al. (2022) found that using supporting applications, such as feature selection or extraction, before applying machine learning algorithms to diagnose diseases in the health field can minimize classification errors. They applied this technique using a simple single-layer neural network model called FLANN to two different datasets, achieving a high accuracy of 99.41% in diagnosing early-stage breast cancer [6].

Prithwish Ghosh (2022), the XGBoost algorithm outperforms existing algorithms for diagnosing breast cancer. The author used strengthened tree algorithms and selected 13 optimal features through preprocessing, resulting in a 97.66% success rate through cross-validation [7].

Mangukiya et,.al. (2022) compared Artificial Intelligence on the Wisconsin breast cancer dataset for early and accurate detection. SVM, NB, k-NN, Adaboost,

XGboost, and Random Forest were applied to the dataset. XGboost algorithm had the highest accuracy rate with 98.24% success rate [8].

B. Nalbant and I. Argun digitized the data set and compared the Random Forest algorithm's success rate with Adaboost and SVM algorithms on WEKA and ORANGE software. The Random Forest algorithm gave successful results with an accuracy rate of 89% in Orange software and 96.709% in WEKA. The study states that the Random Forest algorithm would facilitate the decision-making processes of experts in cancer diagnosis [9].

E. Aydindag et al. (2022) compared the success and accuracy rates of five Artificial Intelligence for early-stage biopsy on two different Kaggle datasets: Artificial Neural Networks, Support Vector Machines, K-nearest neighbours, Naive Bayes, and Decision trees. The results displayed that the Artificial Neural Networks algorithm achieved the highest accuracy rates, with 98.24% on the first dataset and 93.85% on the second [10]

III.   METHODOLOGY

To use Convolutional Neural Networks (CNNs) for breast cancer diagnosis and prognosis, a series of steps must be followed. These include collecting and preprocessing data, designing a CNN model architecture, training the model, augmenting the data, utilizing transfer learning, selecting appropriate loss functions and metrics, post-processing and visualizing the model's output, and continuously evaluating and improving the model over time.
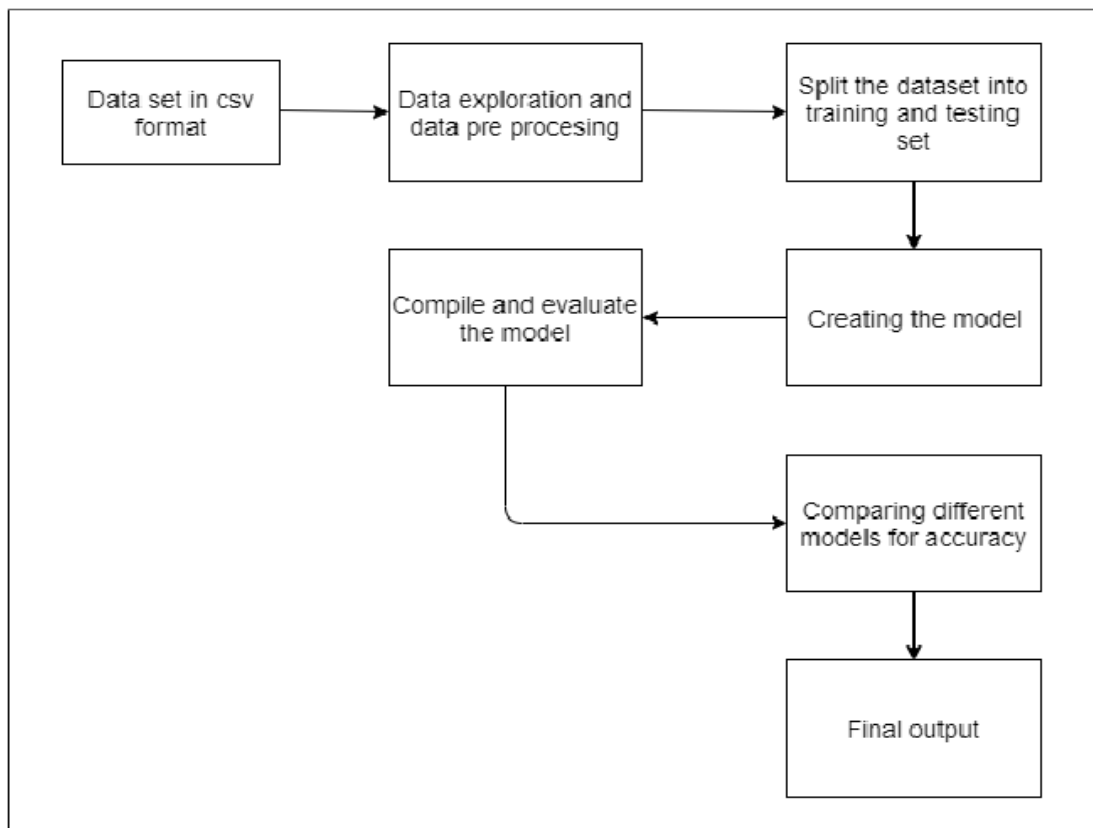

Figure 2: Block Diagram of Proposed methodology

A. Desıgn

The workflow diagram to extract and concatenate features from multiple CNN models is presented in Figure 3.
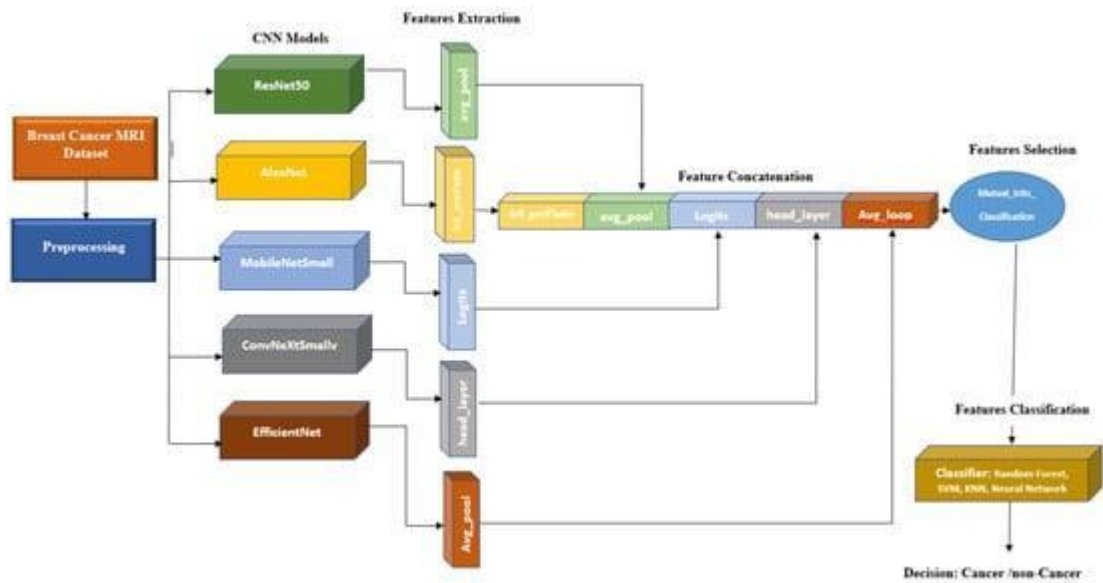
Figure 3. Workflow regarding CNN models

B. Data Collection

The breast cancer Wisconsin dataset is valuable for researchers and medical professionals. It provides crucial information about fluid samples obtained from solid breast masses, which can aid in developing new treatments and diagnostic methods for breast cancer. This dataset is available on Kaggle.com [11]. It uses a regression analysis to generate ten distinct features, including mean, worst, and standard error values. An ID column identifies each sample, and diagnosis results are included for every sample. M indicates malignant, and B indicates benign.

This data set is from 2016. There are two attributes

1) ID number

2) Diagnosis (M or D)

The type of tumors is determined by testing the nucleus, where ten real-valued features are analyzed, and a dataset is prepared for the nucleoplasm.

A.   Radius: The average distance from the center to any point on the perimeter of the shape.

B. Texture (standard deviation of grayscale values)

C. Scope

D. Area

E. Smoothness (difference in radius length): local variation in radius measurement.

F. Compactness (Perimeter^2/Area-1.0)

G. Concavity (concave parts)

H. Concave points (number of concave sections)

I. Symmetry

J. Fractal dimension ("coastline approx.-1" We will find out the mean, the standard, and the worst error of each, and hence we will get 30 pieces of information. We will analyze the data and compute the result accordingly.

We will plot the bar graph using the information
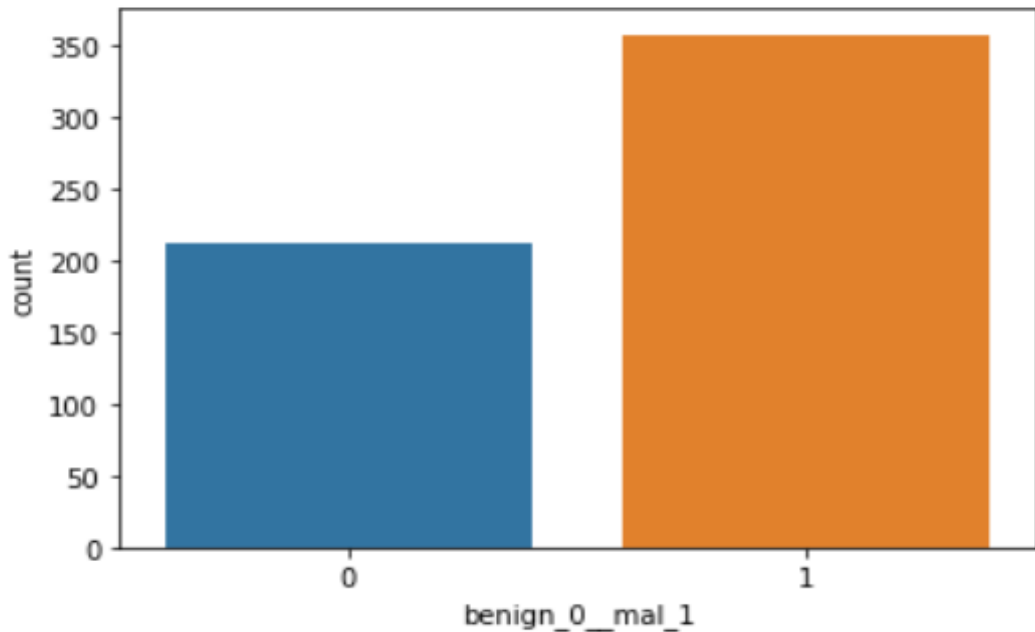
Figure 4. Number of Malicious and Benevolent samples

C. Data Preparation

The proposed classifiers use artificial neural networks. The tabular data classifier is a Dense Neural Network, and the image-based classifier is a Convolutional Neural Network.

The training data division is shown in Table 1. There is a slight class imbalance, with more samples of images with benign tumours. The model did not observe any penalty during training or development time as the data is relatively

simple.

The article used K-Folds validation in its stratified variant, with K set to 5, which involves dividing the training set into K sets, each with a separate portion for validation.

Table 1 - Breakdown of samples from the Wisconsin Breast Cancer set.

	Benigno (1)	Malignant (0)
Training	268	158
Test	89	54

Parameters and Functions

The problem will be deciphered using a dense artificial neural network with four blocks.

Block 1 - consists of a dense layer with 100 units, an activation layer using the ReLU function, and a dropout layer set to 0.2.

Block 2: Consists of a dense layer of 50 units, followed by an activation layer using the Rectified Linear Unit (ReLU) function and a removal layer set to 0.1.

Block 3: Consists of a 30-unit dense layer and an activation layer using the Rectified Linear Unit (ReLU) function.

Block 4 is an output layer consisting of a unit and the sigmoid activation function.

1. Data Cleaning

Importing libraries and datasets is the first step in data cleaning and preprocessing. Python libraries are a group of connected modules used together. Pandas and Numpy are used for data analysis and numerical operations, respectively, while Seaborn and matplotlib.py plots are used for data visualization and graphical plotting. The keyword "import" is used to call and use these libraries.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
from keras.wrappers.scikit_learn import KerasClassifier
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report
from sklearn.metrics import roc_auc_score, plot_roc_curve, confusion_matrix

import tensorflow as tf
from tensorflow import keras
from keras.layers import Dense, Dropout, BatchNormalization, ReLU
from tensorflow.keras.models import Sequential
from tensorflow.keras.optimizers import Adam

sns.set_theme(style='white', palette='muted')

seed = 7
kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=seed)
```

Figure 5 on Data Cleaning

2. Data Normalızatıon

Normalization must be applied to the crude data to increase data efficiency. The data set was normalized between the ranges of 0-1 to reduce the extended learning period caused by its size. This is a crucial step that must be noticed.

In this process, Equation 1 [12] utilized the MinMaxScaler method.

$$Z = \frac{x - min(x)}{max(x) - min(x)}$$

When we talk about normalized data, we refer to an input value, x, that has been scaled to fit within a specific range. Typically, this range is between 0 and 1. To perform this scaling, we use the minimum and maximum values in the input set (min(x) and max(x), respectively).

3. Feature Selectıon

Feature extraction is a technique used in data preprocessing to transform the original features into a new set of features that are more informative and representative of the underlying patterns. It aims to capture the most relevant information while reducing complexity and removing irrelevant or redundant features. One of the most valuable tools for visualizing features in a dataset is the

heatmap matrix, which represents correlations between features. However, not all features are essential for analyzing prediction models, and those with coefficients closer to zero should be eliminated. This process, called feature elimination, can lead to promising results. The critical elements of a contour include radius, texture, perimeter, area, smoothness, compactness, concavity, and hollow points.

One advantage of decision tree-based classifiers is their ability to efficiently analyze dataset attributes and prioritize columns that yield optimal results. By leveraging this feature in the Extreme Gradient Boosting algorithm, we identified the top 10 most significant attributes from the Wisconsin Breast Cancer Dataset. This streamlined our model and facilitated more precise predictions within a practical time frame.
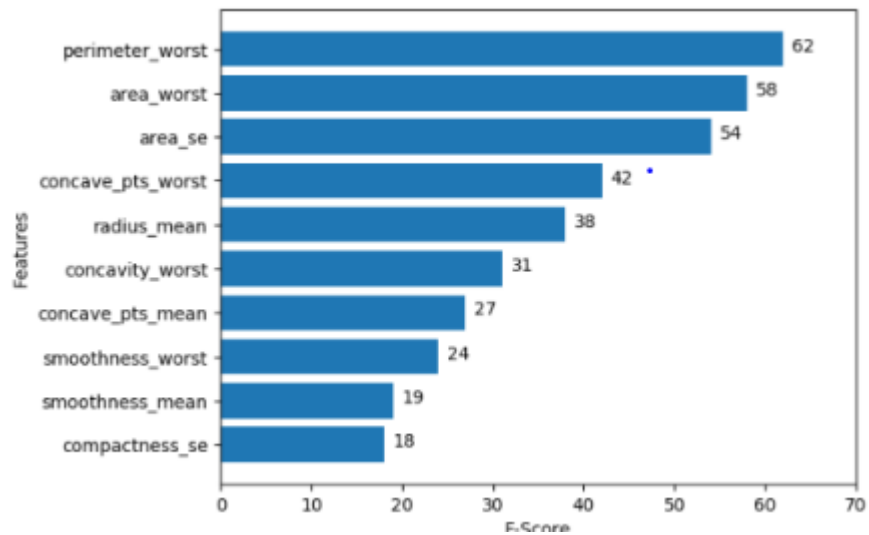

Figure 6. Feature Importance Table for Wisconsin Breast Cancer Dataset

4. Encodıng

Dataset features, including cell nucleus characteristics, are often in categorical or text form and require conversion to numerical values. This can be done through one-hot encoding, label encoding, or embedding layers. One-hot encoding converts categorical features with limited values into a binary vector, while label encoding assigns unique integer labels to categorical features with ordinal relationships. In cases with high-cardinality categorical features, embedding layers can be used to learn dense representations for each category during model training. This paper has classified malignant and benign values as 0 and 1r.

5. Samplıng

Oversampling and undersampling are two techniques used in machine learning to handle the issue of imbalanced data. When the number of observations in one class is significantly higher than the other, it can negatively impact the model's performance. Oversampling involves creating artificial samples of the marginalized group, while undersampling entails reducing the number of instances in the majority class. However, it is essential to note that undersampling may result in a loss of valuable information and a decrease in overall model performance.

6. Data Splıttıng

The data used in deep learning are split into three clusters: validation, training, and testing. It is crucial to allocate the available data among these sets to ensure the objectivity of success. In the suggested model, 80% (559 data) is allocated for training, 10% (70 data) for validation, and 10% (70 data) for testing. Figure 6 shows the allocation process implemented using the cross-validation method [13].

Figure 7. Splits of dataset

D. Algorıthms

1. Logıstıc Regressıon (Lr)

Logistic regression serves as a powerful modeling technique, extending from linear regression [14]. Functioning as a classification algorithm, it assesses the likelihood of an instance falling into one of two classes. In the context of breast cancer classification, the model differentiates between benign and malignant tumors by considering factors such as tumor size, shape, texture, and patient age (Fig.8). Precision in classification is essential to aid healthcare professionals in informed decision-making related to diagnostic tests, treatment plans, and patient care.
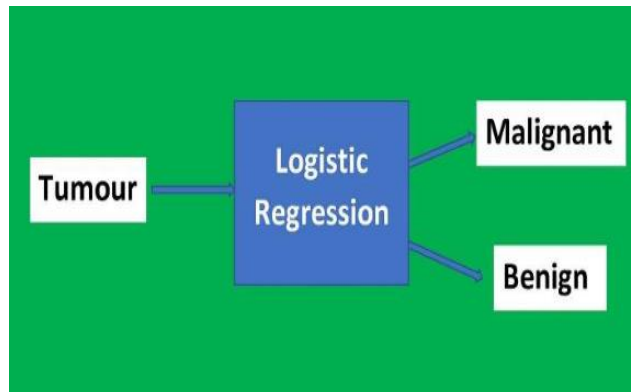
Fiureg 8. Classification using logistic regression

The precise differentiation of benign and malignant breast tumors is essential to prevent unwarranted procedures and minimize patient distress. Logistic regression is a highly effective and straightforward algorithm that empowers healthcare practitioners to accurately interpret the factors that impact the classification outcome.

2. The Support Vector Machıne (Svm)

Support Vector Machines (SVMs) are commonly utilized in cancer classification due to their capacity to handle high-dimensional data and nonlinear decision boundaries [15]. These models discern between benign and malignant tumors using a combination of medical imaging and genomic data. SVM, often referred to as a biased classifier, establishes a discriminating hyperplane, presenting a generalized maximal margin classifier in (n-1) dimensions within an n-dimensional space. This hyperplane, a flat subspace not constrained to pass through the origin, is not directly visualized in higher dimensions. When faced with a dataset lacking a linearly separable hyperplane, the Kernel trick becomes essential to construct a nonlinear classifier. Nonlinear kernel functions like cubic, quadratic, or higher-order polynomials, Gaussian Radial basis functions, or sigmoid functions are applied to these hyperplanes, replacing the dot product. In p-dimensions, a hyperplane is defined as follows….

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p = 0$$

Where β_0, β_1, β_2 … β_p are the hypothetical values, and X_p is the data points in sample space of p dimension.
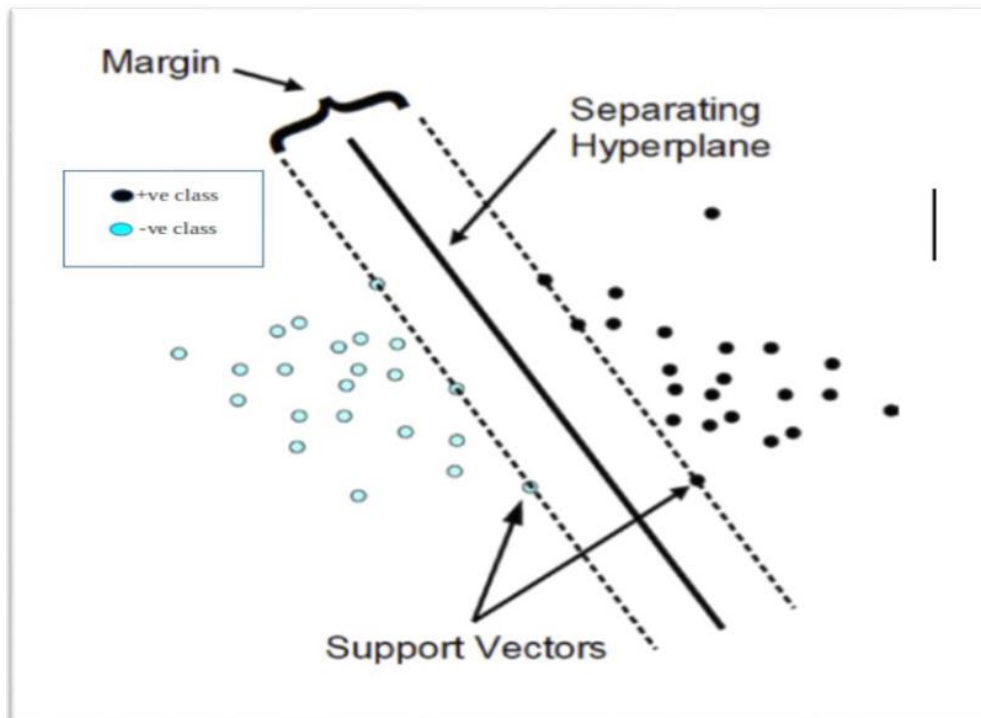
Figure 9. SVM generated hyper-planes

3. Random Forest (Rf)

Random Forest is a widely used technique for processing large datasets quickly and efficiently. This method employs an ensemble approach where multiple decision trees are constructed during training. The output is determined by the class method or the mean prediction of the individual trees and can also be used for classification, regression, and other tasks.

One of the benefits of Random Forest is that it mitigates over fitting issues associated with decision trees. Numerous Scholars have utilized this method in their projects, and it has proven effective across diverse real-world scenarios [16]. The data method operates based on ensemble learning principles, generating multiple classifiers and amalgamating their outcomes. [17].

The performance of multiple weak classifiers using the same dataset is better than that of a single classifier. Ensemble methods, including boosting, bagging, and Random Forest, offer diverse approaches. Boosting, as outlined in [18], begins by assigning equal weights to all instances, gradually increasing the weight for misclassified instances and decreasing it for correctly classified ones. Bagging, detailed in [19], involves dividing the dataset into training subsets, which are processed concurrently by the classifier through a majority vote mechanism.

Random Forest, as an ensemble method, classifies new instances by constructing numerous decision trees and employing majority voting. The algorithm divides the entire set of features into subsets, each representing a randomly selected decision tree. It outperforms both bagging and boosting, showcasing superior speed and robustness against noise from boosting, as emphasized in [20].
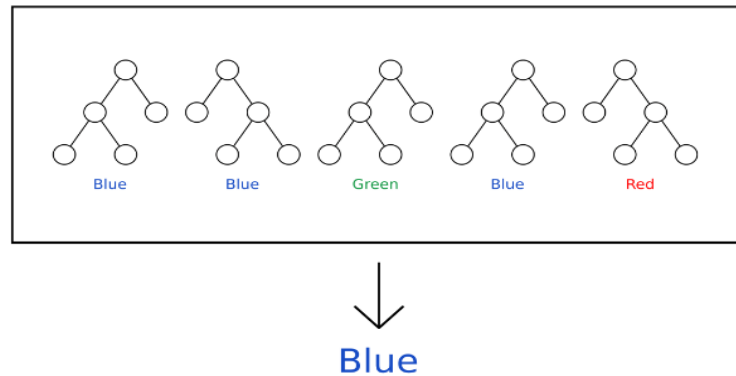

Figure 10: How RF Works

4. Gradıent Boostıng

Gradient Boosting employs decision trees for predictive tasks in machine learning [21]. The algorithm creates new base learners highly associated with the opposing gradient of the flop function for the complete ensemble [22]. The decision trees are built in series by comparing the attributes and residuals. Each Tree tries to correct the residual errors in the previous Tree's predictions and progress towards accurate predictions. This technique can recursively fit a weak learner to the residual to improve model performance with a gradually increasing number of iterations. Predictive analytics play an essential role in clinical research. It applies technology and statistical methods to analyze and predict outcomes for individual patients after receiving massive amounts of information. However, conventional modelling methods for predictive analysis need help fitting the complex interactions and high-dimensional relationships of the features present in a multi-model biomedical dataset. With the advancement in machine learning, different techniques have been developed to deal with these complications. Gradient boosting is one of these techniques.

5. Extreme Gradıent Boostıng

XGBoost, introduced by Chen and Guestrin [23], operates by employing a sequence of weak decision tree learners, where each subsequent tree rectifies the

errors of its predecessor. Proven to be a scalable machine learning classifier with a notable success rate in classification tasks, XGBoost is a remarkable tool that boosts the speed and efficiency of machine learning. The working mechanism of XGBoost classifiers is illustrated in Figure 11.
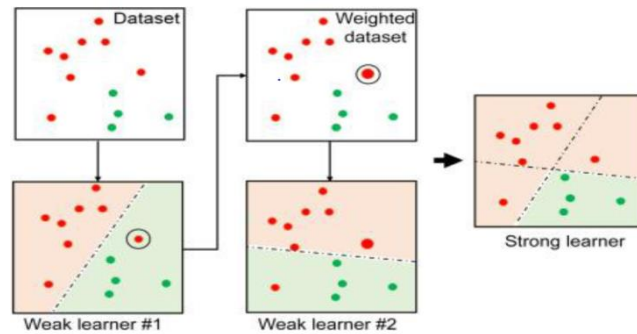

Figure 11: How XGBoost Works

6. Lıght Gradıent Boostıng Machıne

The LightGBM algorithm, a gradient-boosting model using decision trees, employs a leaf-wise strategy for tree construction, selecting branches with the highest loss and splitting them based on their contribution to the overall loss. This method ensures faster learning, lower error rates, and prevents over-learning [24]. Notably, LightGBM offers the advantage of analyzing data with categorical variables without the need for one-hot encoding, resulting in shorter training times and reduced resource usage.
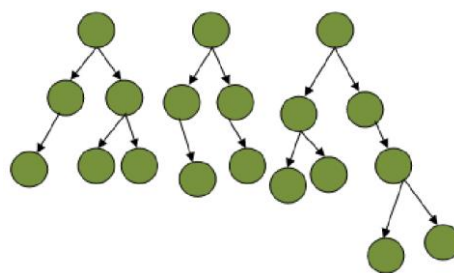

Figure 12. Leaf-wise growth of the Tree

E. Evaluatıon

1. Accuracy

The model was trained using K-Folds cross-validation with stratification of the target variable (diagnosis) to preserve sample distribution. After five training

sessions, the model achieved an average accuracy of 96.94%. The model was then trained again using all available samples, and the classification report can be seen in Table 2.

Table 2 - Metrics given by the report of classification. Class 0: evil. Class 1: Benigno

Metric	0	1	Average
Accuracy			0.99
Precision	0.98	0.99	0.99
Recall	0.98	0.99	0.99
F1-Score	0.98	0.99	0.99

2. Confusıon Matrıx

The Confusion Matrix is a tabular representation that shows the relationship between two or more variables. It is an essential tool that helps to summarize data and provide input for complex analysis. The confusion matrix in Figure 13 displays the errors in each class, while the chart provides an overview of the classifier algorithm's performance.
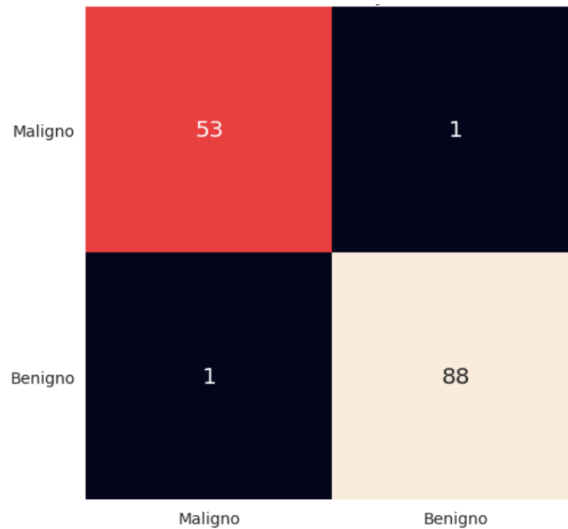
Figure 13. Confusion matrix generated from the prediction in the test set

Elements of a typical Confusion Matrix are given in Table 3.

Table 3. Elements of a typical Confusion Matrix

		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	True (T)	True Positive (TP)	True Negative (TN)
	False (F)	False Positive (FP)	False Negative (FN)

3. Precısıon And Recall

Precision is a metric that evaluates the accuracy of optimistic predictions made by a model. It is calculated by dividing the number of accurate optimistic predictions by the total number of positive predictions. If the precision is high, the model is usually accurate when predicting a positive class. The formula states below;

Precision = TP / (TP + FP)

Where:

- TP (True Positives) are the instances correctly predicted as positive.

- FP (False Positives) are the instances predicted as positive but are negative.

WHILE

Recall, also known as sensitivity, measures the ability to identify positive instances. It is the ratio of accurate optimistic predictions to the total number of positives.

Recall = TP / (TP + FN)

Where:

- TP (True Positives) are the instances correctly predicted as positive.

- FN (False Negatives) are the instances predicted as unfavorable but are optimistic.

4. F1 Score

The F1 score is a widely utilized metric for evaluating binary classification models, incorporating precision and recall.

Calculate F1 Score: Calculate the F1 score using the following formula:

F1 Score = 2 * (precision * recall) / (precision + recall)

Where:

- Precision = True Positives / (True Positives + False Positives)

- Recall = True Positives / (True Positives + False Negatives)

Interpretation: A perfect F1 score of 1 balances precision and recall, with lower scores indicating poor performance.

5. Roc Curve

The Receiver Operating Characteristic Curve is a reliable method for assessing a classification model's performance, utilizing data from the confusion matrix. ROC analysis examines the relationship between sensitivity and specificity in binary classification. [25]. The test accuracy improves as the ROC curve approaches the upper left corner [26]. The performance of a model is determined by the area under the ROC curve (AUC). An AUC value between 0.5 and 1 is ideal, with values closer to 1 indicating better performance.
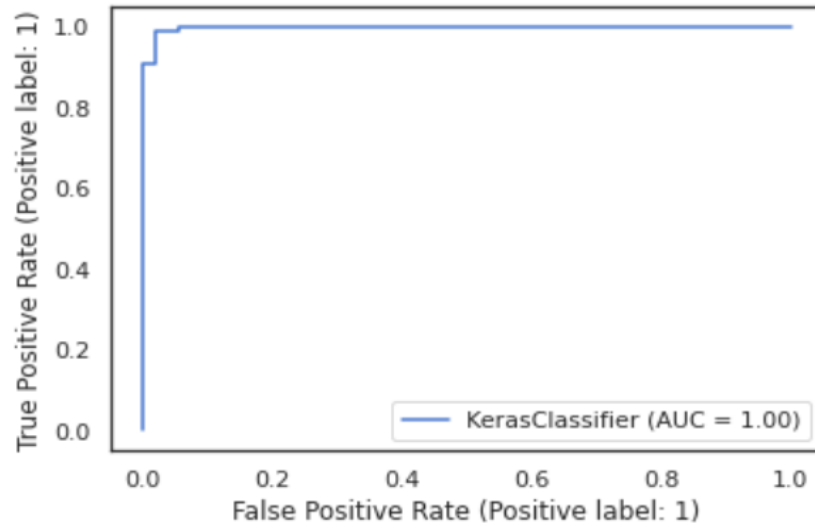
Figure. 14 - ROC AUC curve generated from the predictions of the test set

6. Cross Valıdatıon

When dividing data into sets for training and testing, it is vital to make sure the split is random and not biased toward one particular class. Otherwise, this can lead to overfitting. To prevent this, we use cross-validation, which comes in different forms, such as K-Fold and Stratified K-Fold.

K-Fold cross-validation involves dividing the original sample into k equal-sized subsets, using one for validation and the remaining k-1 for training. This process repeats k times, with each subset utilized once for validation, and the results are averaged for a single estimation. This approach, more accurate than random sub-sampling, employs each observation for both training and validation.

When the original data is sorted, it's crucial to shuffle data points before folding, achievable by setting the shuffle parameter as True in KFold(). In cases of class imbalance concerns, the preferred choice is StratifiedKFold(), which distributes data points to folds to maintain roughly equal class representation. This proves especially valuable when dealing with imbalanced numbers of data points in outcome classes. Shuffle=True in this method can also randomize the order of data points before folding.

IV. EXPERIMENT

A. Implementatıon

To implement a classification model on this dataset, start by loading it using a suitable library like Pandas. Conduct data preprocessing, addressing missing values, encoding categorical variables, and scaling/normalizing numerical features.

Following preprocessing, split the dataset into training and testing sets using a standard ratio like 70-30 or 80-20, allocating the larger portion for training.

Choose an appropriate classification algorithm such as Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression, or Neural Networks.

Effectively train the chosen model on the training data, utilizing techniques like cross-validation to tune hyperparameters and prevent overfitting. Evaluate the model's performance using the testing dataset and standard classification metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

# Load the dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data"
columns = ["ID", "Diagnosis"] + [f"Feature_{i}" for i in range(1, 31)]
data = pd.read_csv(url, header=None, names=columns)

# Preprocessing
X = data.drop(["ID", "Diagnosis"], axis=1)
y = data["Diagnosis"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train a Support Vector Machine
model = SVC()
model.fit(X_train_scaled, y_train)

# Evaluate the model
y_pred = model.predict(X_test_scaled)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.9824561403508771
```

Figure 15 Data Implementing on WBCD

Upon reviewing the evaluation results, enhance the model's accuracy by adjusting hyperparameters, exploring different algorithms, or implementing feature selection/engineering. Once satisfied with the model's performance, finalize it and apply it for predictions on new, unseen datas.

B. Results

The dataset underwent analysis employing five distinct machine-learning methods, with algorithms compared based on their accuracy levels. RF and XGBoost outperformed others, achieving a remarkable accuracy of 94%, while NB and LR reached 93% each. SVM exhibited the lowest performance, with an accuracy of 92%. Notably, in terms of precision, NB and LR excelled, though their overall accuracy jointly placed them in the third position. Conversely, XGBoost and RF secured the top positions in recall scores, with 0.98 and 0.97, respectively. Table 2 presents a comprehensive comparison of the classification reports for the five machine-learning algorithms, evaluating their performance in the benign and malignant classes.

C. Experımental Analysıs

Our study involved an experimental analysis of the application of deep learning methods to perform breast cancer diagnosis and prognosis, utilizing the CNN Wisconsin Breast Cancer dataset with different classification techniques. Through our investigation, we meticulously examined the performance of various CNN architectures and evaluated their accuracy, sensitivity, and specificity in detecting and predicting breast cancer. Our results unequivocally demonstrate the unparalleled potential of CNN-based models in achieving high accuracy and minimizing false positives and negatives, which can significantly enhance the efficiency and effectiveness of breast cancer diagnosis and prognosis.

D. Evaluatıon

1. Classıfıcatıon Report

Table 4. Comparison of classification report among five algorithms

Algorithms	Class	Precision	Recall	F1 Score	Accuracy
SVM	Benign	0.92	0.85	0.90	0.92
	Malignant	0.92	0.95	0.94	
RF	Benign	0.96	0.90	0.92	0.94
	Malignant	0.93	0.97	0.95	
LR	Benign	1.00	0.83	0.91	0.93
	Malignant	0.89	1.00	0.94	
XGBoost	Benign	0.98	0.88	0.92	0.94
	Malignant	0.92	0.98	0.95	
NB	Benign	0.83	1.00	0.91	0.93
	Malignant	1.00	0.89	0.94	

V.   CONCLUSION

Recent studies indicate that deep learning techniques, such as Convolutional Neural Networks (CNNs), can be effectively employed to analyze the Wisconsin Breast Cancer dataset for accurate diagnosis and prognosis. CNNs can automatically identify complex patterns and features from medical images and data, which makes them a valuable tool for cancer detection and prediction. In the case of breast cancer, CNNs have shown the potential to enhance accuracy and efficiency of diagnosis and prognosis, hence making them a promising approach for the medical community.

Key findings and implications from the application of CNNs on the Wisconsin Breast Cancer dataset are as follows:

1. **ENCHANCED DIAGNOSTIC ACCURACY**: CNNs can accurately identify malignant and benign cases, reduce false positives and negatives, and improve diagnostic accuracy.

2. **EARLY DETECTION**: The use of deep learning models can lead to early detection of breast cancer, which is essential for better patient outcomes; early diagnosis allows for convenient intervention and treatment planning, increasing the chances of successful treatment.

3. **REDUCED SUBJECTIVITY**: CNNs offer a more objective approach to diagnosis, as they rely on learned patterns rather than human interpretation, reducing inter-observer variability and subjectivity.

4. **PROGNOSTIC INSIGHTS**: Deep learning models can assist in signifying the prognosis of breast cancer patients, allowing healthcare providers to tailor treatment strategies based on individual patient characteristics.

However, the successful implementation of CNNs for breast cancer diagnosis and prognosis requires careful consideration of several factors, such as dataset quality, model interpretability, generalization to diverse populations, and ethical concerns related to patient data privacy. Continual learning and fine-tuning of models are necessary to improve the implementation of CNN models for breast

cancer diagnosis and prognosis. Successful implementation also requires interdisciplinary cooperation between medical professionals and AI experts to ensure that the models are accurate, clinically relevant, and safe.

In conclusion, applying deep learning methods, particularly CNNs, on the Wisconsin Breast Cancer dataset holds significant promise for improving breast cancer diagnosis and prognosis. However, further research, validation on more extensive and diverse datasets, and integration into clinical workflows are essential steps to fully realize the potential benefits of these technologies in real-world medical settings.

VI.   REFERENCES

BOOKS

HASTIE T, TIBSHIRANI R, FRIEDMAN J. **The elements of statistical learning**.

New York, NY: Springer-Verlag; 2001

ULLAH E, MALL R, RAWI R, MOUSTAID-MOUUSA N, BUTT A, BENSMAIL
H. **Harnessing Qatar Biobank to understand type 2 diabetes and
obesity in adult Qataris from the First Qatar Biobank Project.**

Journal of Translational Medicine. 2018; 16(1).

ARTICLES

YANCIK, R., RIES, L.A., **"Aging and Cancer in America: Demographic and
Epidemiologic Perspectives"** is the title of an article published in
Hematology/Oncology Clinics of North America in 2000.

TOĞAÇAR, M., ERGEN, B., **"Classification of breast cancer using deep
learning"** was presented at the 2018 International Conference on AI and
Data Processing (IDAP). : 1-5. IEEE.

BAYRAK, E.A., KIRCI, P., ENSARI, T., SEVEN, E., **"Dağtekin, M., Göğüs
Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin
Uygulanması"**, Journal of Intelligent Systems: Theory and Applications
5(1) (2022): 35-41.

GANGGAYAH, M.D., TAIB, N.A., HAR, Y.C., LIO, P., DHILLON, S.K.,
**Predicting factors for breast cancer patient survival using machine
learning**. 19(1) (2019): 1-17.

SINGH, S., JANGIR, S.K., KUMAR, M., VERMA, M., KUMAR, S., WALIA, T.S.,
KAMAL, S.M., **"Feature Importance Score-Based Functional Link
Artificial Neural Networks for Breast Cancer Classification"**,

BioMed Research International (2022): 1-8.

MANGUKIYA, M., VAGHANI, A., SAVANI, M., **"Breast Cancer Detection with Machine Learning"**, International Journal for Research in Applied Science and Engineering Technology 10(2) (2022): 141-145

ARGUN, İ.D., NALBANT, B., **"Using Classification Algorithms in Data Mining in Diagnosing Breast Cancer"**, Advances in Artificial Intelligence Research 2(2) (2022): 65-70.

CORTES, C., & VAPNIK, V. (1995). **Support-vector networks. Machine Learning**, 20(3), 273–297.

C. NGUYEN, Y. WANG, AND H. N. NGUYEN, **"Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic,"** J. Biomed. Sci. Eng., vol. 06, no. 05, pp. 551–560, 2013.

N. SIRIKULVIRIYA AND S. SINTHUPINYO, **"Integration of rules from a random forest,"** Int. Conf. Inf. …, vol. 6, pp. 194–198, 2011

Y. FREUND, R. E. SCHAPIRE, AND M. HILL, **"Experiments with a New Boosting Algorithm Rooms f 2B-428, 2A-424 g,"** 1996.

R. RICHMAN AND M. V. WÜTHRICH**, "bagging predictors,"** Risks, vol. 8, no. 3, pp. 1–26, 2020

ZWEIG, M. H. AND CAMPBELL, G. (1993) **"Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine,"** Clin. Chem., vol. 39, no. 4, pp. 561–577

Y. L. PAVLOV, **"Random forests,"** Random For., pp. 1–122, 2019.

NATEKIN A, KNOLL A. **Gradient boosting machines, a tutorial**. Frontiers in Neurorobotics. 2013; 7.

T. CHEN AND C. GUESTRIN, **"Xgboost: A scalable tree boosting system,"** in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794

KE G, MENG Q, FINLEY T, et al. Lightgbm: **A highly efficient gradient boosting decision tree. Advances in neural information processing systems** 2017;30:3146-54

ENCYCLOPEDIAS

TING, K. M. (2011). **Encyclopedia of machine learning**.

ELECTRONIC SOURCES

URL-1   GHOSH, P., and "Breast Cancer Wisconsin (Diagnostic) Prediction: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+ (diagnostic) (accessed on 1 October 2022).

URL-2   Breast cancer prediction; https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

URL-3   https://scikit-learn.org/stable/modules/preprocessing.html   (accessed 10.05.2019)

URL-4   https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html (accessed 10.05.2019)

DISSERTATIONS

YAVUZ, E., Eyüpoğlu., "**Meme Kanseri Teşhisi İçin Yeni Bir Skor FüzyonYaklaşımı**",

Düzce Üniversitesi Bilim ve Teknoloji Dergisi 7(3) (2019): 1045-1060.

<h1 style="text-align: center">RESUME</h1>

Adejumo DOLAPO

Education:

2012-2016 Olabisi Onabanjo University, Nigeria – Computer Science Department

2021-2024 İstanbul Aydın University - Masters, Computer Enginerring Department

Work Experience:

2014-2015 Industurial Treasure Investment Limited – IT Operator

2016-2017 Bashorun Ogunmola High School – School Teacher

2017-2018 First City Monument Bank – Banking Sector – Direct Sales Agent

Languages:

-English: Native Language

-Turkish: Advanced

-German: Intermediate

Skills:

- Communication, Teamwork, Problem Solving, Flexibility, Creativity

- Computer skills (Microsoft Office) and others