Advances in Artificial Intelligence and Machine Learning; Research 3 (3) 1494-1525 Received 22-07-2023; Accepted 23-09-2023; Published 30-09-2023

Feature Selection and Comparative Analysis of Breast Cancer Prediction Using Clinical Data and Histopathological Whole Slide Images

Sarfaraz Ahmed Mohammed mohammsm@mail.uc.edu
Department of Computer Science
University of Cincinnati
Cincinnati, OH 45221-0030, USA

Senuka Abeysinghe senuka.abeysinghe24@ishd.us
Indian Hill High School
Ohio's College Credit Plus Program
Cincinnati, OH 45243, USA

Anca Ralescu ralescal@ucmail.uc.edu
Department of Computer Science
University of Cincinnati
Cincinnati, OH 45221-0030, USA

Corresponding Author: Sarfaraz Ahmed Mohammed

Abstract

Breast Carcinoma is a common cancer among women, with invasive ductal carcinoma and lobular carcinoma being the two most frequent types. Early detection is critical to prevent cancer from becoming malignant. Diagnostic tests include mammogram, ultrasound, MRI, or biopsy. Machine Learning algorithms can play a key role in analyzing complex clinical datasets to predict disease outcomes. This study uses machine learning and deep learning techniques to analyze publicly available clinical and medical image data. For clinical data, Principal Component Analysis (PCA) and Particle Swarm Optimization (PSO) are applied on the Wisconsin Breast Cancer dataset (WDBC) for feature selection and evaluate the performance of each modality in distinguishing between benign and malignant tumors. The results obtained show that the Random Forest (RF) classifier outperforms other classification algorithms using both PSO and PCA feature selections, achieving predictive accuracies of 95.7% and 97.2% respectively. The first part of the paper contains a comprehensive analysis of the two feature selection methods on clinical data to optimize predictive performance. The second part of the paper is concerned with image data. Although Histopathological Whole Slide Imaging (WSI) has been validated for a variety of pathological applications for over two decades of manual detection of cancerous tumors, it remains challenging and prone to human error. With the potential of deep learning models to aid pathologists in detecting cancer subtypes, and the increasing predictive ability of current image analysis techniques in identifying the underlying genomic data and cancer-causing mutations, the second half of the paper focusses on feature extraction using a deep convolutional neural network (U-Net)

Citation: Sarfaraz Ahmed Mohammed et al. Feature Selection and Comparative Analysis of Breast Cancer Prediction Using Clinical Data and Histopathological Whole Slide Images. Advances in Artificial Intelligence and Machine Learning. 2023;3(3):87.

trained on WSI's from The Cancer Genome Atlas (TCGA) to accurately classify and extract relevant features. The focus is on feature extraction, nuclei-based instance segmentation, H&E-stained image extraction, and quantifying intensity information for a given WSI to classify the disease type. A comprehensive analysis of feature selection methods is presented for both clinical and medical image data.

Keywords: Breast cancer, Machine learning, Principal component analysis, Particle swarm optimization, Feature selection, Logistic regression, Naïve bayes classification, k-NN, Support vector machines, Random forest, K-Means, Whole slide images, TCGA, Histopathology, Deep learning, Digital image analysis, Convolutional neural network, H&E-stained images, Nuclei segmentation.

1. INTRODUCTION

Breast Carcinoma is a major health concern worldwide and the second leading cause of death in women in the United States. According to the CDC, approximately 264,000 cases are diagnosed in women and 2,400 cases in men annually [1]. The World Health Organization (WHO) reports that in 2020, 2.3 million cases were diagnosed among women, resulting in 685,000 deaths worldwide, making it the most prevalent cancer [2]. Various risk factors have been identified for breast cancer, including age (over 50), dietary habits, heredity, reproductive history, alcohol consumption, being overweight, and hormone replacement therapy, among others.

In recent years, data mining has become an important tool for discovering hidden patterns and extracting useful information from large datasets. Feature selection is a first step in data preprocessing, as it helps to identify the most relevant features for building efficient machine learning models. Dimensionality reduction techniques such as PCA, have been widely used to reduce the number of features while preserving information [3], leading to increased predictive accuracy and learning efficiency. The first half of this paper compares the performance of five popular classification algorithms - Support Vector Machines [4], Naïve Bayes [5], K-Nearest Neighbors [6], Logistic Regression [7], and Random Forest [8] - on the WDBC dataset. PCA and PSO are employed as a dimensionality reduction technique for feature selection. It is inferred that PSO has shown promising results in extracting a proper subset of features [9–11]. Compared to other evolutionary algorithms PSO is computationally less expensive and converges more quickly [12]. However, the results obtained in this paper show that PCA results in better classification accuracy and performance than PSO and in each of the two feature selections, the Random Forest classifier outperforms the others in terms of accuracy in diagnosing on the WDBC dataset.

Histopathological image analysis on whole slide images (WSIs) has seen significant advancements in recent years, thanks to the use of various deep learning models in the field of computational pathology. These models aid pathologists in screening image samples at the highest resolution and detecting cancer patterns, such as differentiating between normal and abnormal tissue segmentation [13], predicting cancer stage [14], and survival rate [15], among other applications. Given the abundance of growing WSIs, it is imperative to analyze them using deep learning techniques that break down the images into patches and perform patch-based optimization to train a whole slide gigapixel image. The second half of this paper outlines various techniques for feature extraction in WSIs and for quantifying the intensity of cell information.

From this point, the paper is organized as follows: Section 2 provides an overview of related work found in the literature on both clinical data and histopathological image analysis. Section 3 describes the feature selection techniques such as Principal Component Analysis (PCA), and Particle Swarm Optimization (PSO) to evaluate and identify the top features obtained from both PCA and PSO and compare the performance on five widely used supervised classification algorithms. Section 4 discusses feature extraction techniques in whole slide images (WSI's) and how to quantify the intensity information. Finally, Section 5, presents conclusions and outlines potential avenues for future research.

2. BACKGROUND STUDY AND RELATED WORK

Medical data is often very complex and large, making it difficult to analyze manually. Data mining techniques, such as classification, can help to discover hidden patterns in the data that may be difficult to identify otherwise. By using these techniques, medical researchers can analyze large amounts of data quickly and accurately and identify important patterns and trends. Classification is a supervised machine learning technique that is often used in medical data analysis to predict a particular outcome, such as the presence of breast tumor as benign (non-cancerous) or malignant (cancerous). There have been many successful applications of classification techniques in medical data analysis, including breast cancer diagnosis. These techniques have been used to develop predictive models that can accurately classify breast cancer cases based on a range of different factors, including patient age, family history, and biopsy results. This section reviews the background studies carried out on both clinical data (WDBC) and histopathological whole slide images (WSIs) from TCGA.

2.1 Related Work on Clinical Data

The performance of various classification algorithms including Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k-Nearest Neighbors (k-NN), using the Wisconsin Diagnosis Breast Cancer (WDBC) dataset was studied extensively [16]. According to reported experimental results, the SVM classifier achieved the highest accuracy of 97.13%. In [17], the authors compared the performance of various classification algorithms, including Naïve Bayes, SVM, Radial Basis Neural Networks (RBNN), Decision Tree, and CART, to identify the best classifier for the WDBC dataset. According to their experimental results, the SVM classifier achieved a high accuracy of 96.99% for both binary and multiclass classification. In [18], the authors applied multilayer perceptron, k-Nearest Neighbor, genetic programming, and random forest algorithms to classify the disease category of the WDBC dataset as either Benign or Malignant. According to their experimental results, the random forest classifier outperformed the other classification algorithms by achieving an accuracy of 96.24%. In [19], the authors proposed a decision tree classifier to predict 5-year survivability of breast cancer based on imbalanced data. To evaluate their models, the authors incorporated a Bagging technique that supports decision tree models to predict the disease outcome. In [20], the authors illustrated and compared various machine learning algorithms on the WDBC dataset, which contains features from digitized images based on tests carried out on breast masses. According to the experimental results, the ML algorithms performed well, with each of them achieving over 90% test accuracy on the classification task. In [21], the authors proposed a

hybrid model consisting of Random Forest and logistic regression classification techniques on the WDBC dataset. The model utilizes top-k features from the dataset as inputs to the logistic regression classifier to analyze and predict breast cancer survivability. TABLE 1. displays the accuracies found in the literature and compares these with our proposed approach.

Table 1: Accuracy Comparison with other studies

Study	**Title**	**Technique**	**Accuracy**	**Accuracy with top features (Our Approach)**
S. Aruna et al., 2011	Knowledge based Analysis of various statistical tools in detecting breast cancer.	Naïve Bayes, Support Vector Machine (SVM), Radial Basis Neural Networks (RBNN), Decision Tree and CART and finds the best classifier for the WDBC dataset.	96.99% (SVM classifier)	97.20% (RF classifier with 12 features using PCA) 95.70% (RF classifier with 12 features using PSO)
Hiba Asri et al., 2016	Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis	Naïve Bayes, Support Vector Machine (SVM), Radial Basis Neural Networks (RBNN), Decision Tree and CART and finds the best classifier for the WDBC dataset.	97.13% (SVM classifier)	97.20% (RF classifier with 12 features using PCA) 95.70% (RF classifier with 12 features using PSO)
Abien Fred Agarap, 2018	On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic dataset	Different machine learning algorithms are used	Approx 90%	97.20% (RF classifier with 12 features using PCA) 95.70% (RF classifier with 12 features using PSO)
Arpit Bhardwaj et al., 2022	Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification	Multilayer perceptron, k-Nearest Neighbor, genetic programming, and random forest on the WDBC dataset to classify the disease.	96.99% (RF classifier)	97.20% (RF classifier with 12 features using PCA) 95.70% (RF classifier with 12 features using PSO)

2.2 Related Work on Histopathological Image Analysis

For a given WSI, some of the problems in computational pathology such as predicting the cancer grade, identifying tumor regions, predicting the survival rate, response to certain treatments, identifying biomarkers to name a few bring several interesting challenges as WSIs are multi-resolution and constitute of a gigapixel image data per slide as these glass slides are captured through whole slide scanners that produces a higher throughput and higher resolution images. These WSIs are too large for a convolutional neural network to handle and require approximately 1000 GB GPU. Moreover, detailed annotations for various tumor regions and hence detailed level of WSI labeling, are difficult to obtain from a pathologist. The learning problem in this case can be regarded as a weak supervision problem. In [22], the authors propose a deep multiple-instance learning frame- work (MIL) where each whole slide is regarded as a bag full of patches obtained from one WSI. It can be inferred from [22], that the framework introduced is a classification problem where each WSI bag (containing extracted patches) is categorized as either benign or malignant with a property that if there exists even a single patch containing malignant cells, then the complete image is considered as Malignant. The authors evaluated the framework on three different datasets and show that BreakHis achieves the highest accuracy in comparison to the other two datasets. However, the issue with the MIL framework is that patch representations do not always capture enough information as the extracted features may span multiple patches. However, in [23], the authors propose an alternative way of modeling the WSIs using a graph convolutional neural networks based (GNN) model. As WSIs are divided into patches and because of their large size and memory constraints, the risk of losing the visual context limits the understanding of the cellular architecture.

2.3 Main Contributions of the Paper

1. Two feature selection methods, namely Principal Component Analysis (PCA) and Particle Swarm Optimization (PSO) have been used on the WDBC clinical dataset with the goal to identify the top features and compare the performance across five of the most widely used supervised classification algorithms, including Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machines, and Random Forest. The study yields significant insights into the effectiveness and efficiency of various classification algorithms for predicting breast cancer type. It emphasizes the critical role of feature selection techniques in enhancing classification accuracy. Apart from this, the paper also throws light on the interpretability of clinical data using k-Means to see the percentage of benign and malignant cells in each of the k clusters.

2. As part of the medical image analysis, we explore feature extraction in WSI. To that end, StarDist is used with a deep convolutional neural network (U-Net) as a backbone model trained on WSI's. Concepts such as H&E staining, nuclei-based instance segmentation as part of feature extraction are carried out to quantify the intensity information for a given WSI.

3. EXPERIMENTAL EVALUATION OF FEATURE SELECTION USING PCA AND PSO

Feature Selection is a challenging task as it involves searching through a large space for possibly interacting features. With the increase of the dimensionality of feature space the amount of data required for making good decisions or achieving good decision boundaries increases exponentially. There are three main methods for feature selection, namely, the Wrapper method [24], the Filter method [25], and the Hybrid method [26]. However, the first two methods can be computationally expensive for high dimensional feature spaces. To address this challenge, a Hybrid model that combines the Wrapper and Filter methods has been developed, which is effective in handling high-dimensional data. Feature dependency plays a crucial role in feature selection, and various measures are available for it in the literature. One of the most popular measures is Correlation-based feature selection [27], based on the correlation coefficient of pairs of features.

Correlation-based feature selection is a technique that involves measuring the similarity between two features based on the correlation coefficient between them. FIGURE 1. illustrates the architecture of the proposed approach.

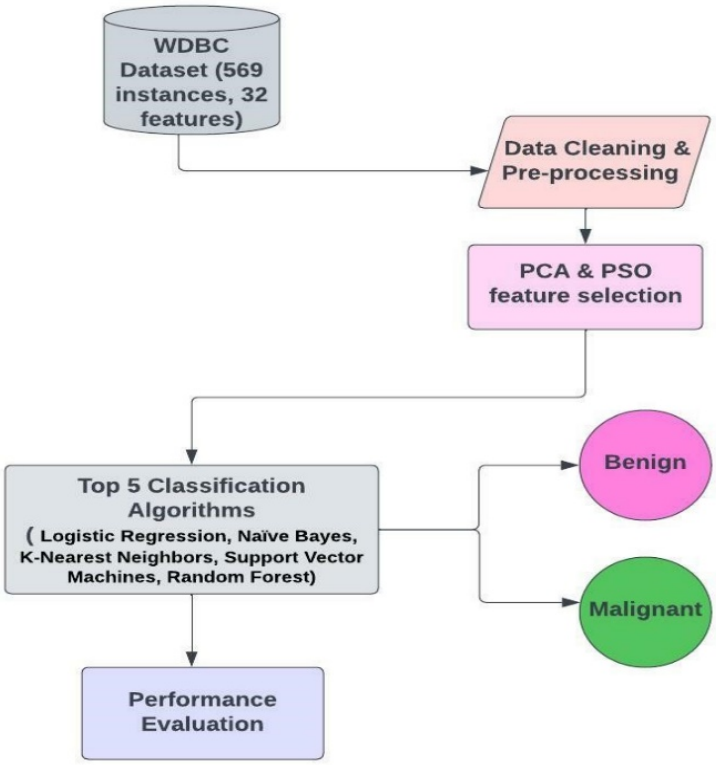


Figure 1: Architecture of the Proposed Approach

3.1 Principal Component Analysis

PCA is a technique used to reduce the dimensionality of datasets while minimizing information loss and increasing interpretability. Interpretability can be increased by generating new variables that are uncorrelated and that can maximize the variance. The goal is to find new variables, called principal components, that maximize variance and are uncorrelated with each other. These principal components are obtained by solving eigenvalue/eigenvector problems [28], which reduces the original dataset. Additionally, PCA addresses the issue of multicollinearity, as all principal components are orthogonal to one another. By preserving as much variability (statistical information) as possible, PCA makes it easier to analyze and understand complex datasets.

PCA consists of several steps as follows: (1) The data is standardized (centered) by finding the mean of all feature values and calculating how far each data point is from the mean. These new values become the centered data D; (2) Obtain a covariance matrix A by finding the variance of each feature column and calculating the pairwise covariance between the features. Positive and negative covariances represent an increase/decrease in the two features or features in opposite directions, respectively [29]. This means that the positive covariance between the two features indicate that the features tend to move together (either increase or decrease), whereas negative covariance indicate that these features move in the opposite directions; (3) Find the eigen values of the covariance matrix using $| A - \lambda I | = 0$. Since there are 30 features, this step yields 30 eigen values, represented as $\lambda 1, \lambda 2, \ldots, \lambda 30$; (4) Calculate the corresponding eigen vectors by normalizing them to unit length and sorting them from highest to lowest eigen values. The direction of the maximum variance, or principal components, is determined by the eigenvectors of the covariance matrix, while the magnitude is defined by the corresponding eigen values; (5) Select the most important eigen values and eigen vectors by ranking the corresponding eigen vectors based on the decreasing order of the eigen values; (6) Finally, a new matrix of the transformed data, called the projection matrix of the important eigen vectors, is obtained by multiplying the centered data D from step 1 with the eigen vectors. This transformed data is represented as the Principal Component scores, such as PC1, PC2, …, PC30, seen in TABLE 2, and it represents the original centered data in the principal component space.

Table 2: Summary of the PCA object

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	3.6444	2.3857	1.67867	1.40735	1.28403	1.09880	0.82172	0.69037	0.6457	0.59219	0.5421
Proportion of Variance	0.4427	0.1897	0.09393	0.06602	0.05496	0.04025	0.02251	0.01589	0.0139	0.01169	0.0098
Cumulative Proportion	0.4427	0.6324	0.72636	0.79239	0.84734	0.88759	0.91010	0.92598	0.9399	0.95157	0.9614
	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22
Standard deviation	0.51104	0.49128	0.39624	0.30681	0.28260	0.24372	0.22939	0.22244	0.17652	0.1731	0.16565
Proportion of Variance	0.00871	0.00805	0.00523	0.00314	0.00266	0.00198	0.00175	0.00165	0.00104	0.0010	0.00091
Cumulative Proportion	0.97007	0.97812	0.98335	0.98649	0.98915	0.99113	0.99288	0.99453	0.99557	0.9966	0.99749
	PC23	PC24	PC25	PC26	PC27	PC28	PC29	PC30			
Standard deviation	0.15602	0.1344	0.12442	0.09043	0.08307	0.03987	0.02736	0.01153			
Proportion of Variance	0.00081	0.0006	0.00052	0.00027	0.00023	0.00005	0.00002	0.00000			

TABLE 2, shows the 30 principal components (PC1-PC30) along with the total variation in the WDBC dataset. It can be inferred that PC1 extracts 44.3% of the total variance and PC2 extracts 19% of the total variance (see FIGURE 2). This means that PC1 (the first principal component) spans the direction of the most variation in the data, while PC2 (the second principal component) spans the direction of the next variation. Together, PC1 and PC2 explain 63% of the total variance. Therefore, by identifying the position of a sample with respect to PC1 and PC2, a reasonable inference can be made in conjunction with the other samples. FIGURE 2 (i) is a biplot [30], of the correlation matrix (PCA) that shows the position of each sample in terms of PC1 and PC2 using the 'ggbiplot' package provided in R, which provides a better visualization of how samples relate to each other. Each principal component is one-dimensional and has a midpoint of 0. The direction that a given variable moves in each PC on a single dimension vector can be either positive or negative. The feature markers are represented by arrows and id by numbers. The tightly clustered feature markers in the biplot are considered highly correlated features.

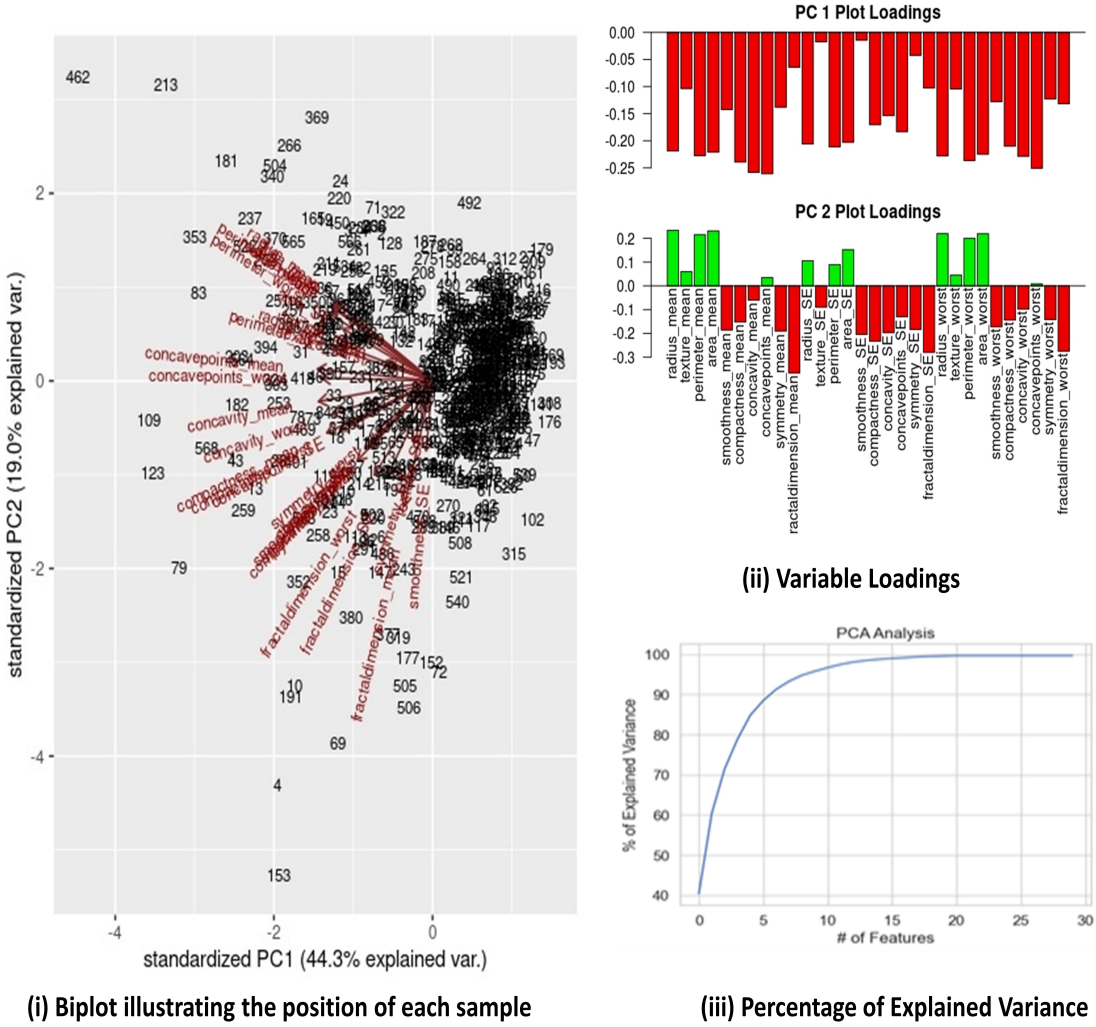


Figure 2: Biplot illustrating the position of each sample in terms of PC1 and PC2

It can be inferred from TABLE 2, that the number of principal components is the same as the number of features (variables), indicating that the number of features isn't reduced. However, since PC1 and PC2 capture around 63% of the total variance, they can be interpreted as storing almost all the information of the data. Hence, the remaining PCs (PC3-PC30) can be ignored as they do not contain significant information.

FIGURE 2 (ii) presents a bar plot of the variable loadings obtained from the PCA, which enables the determination of variables that positively and negatively impact the principal components, PC1 and PC2. Notably, the smoothness_SE is the most influential variable in PC1, while the concavepoints_mean is the least influential. Similarly, the radius_mean is the most influential variable in PC2, while the fractaldimension_mean is the least influential. It is worth noting that PCA possesses certain undesirable features in situations where variables have different units of measurement. PCA relies on a variance criterion that is affected by varied measurement units, and the principal components of some covariance matrices can change if there exist some measurement units on one or more variables. To address this undesirable feature, it is crucial to standardize the variables. FIGURE 2 (iii) displays the percentage of explained variance, indicating that by incorporating 12 features, approximately 97% of the total variance of the data can be preserved, thereby justifying the implementation of PCA for these 12 best features.

3.1.1 Experimental evaluation

For this study, we utilized the Wisconsin Diagnosis Breast Cancer (WDBC) dataset, which comprises 569 instances with 32 columns. These columns represent features extracted from a digitized image of a fine needle aspirate (FNA) of a breast mass, detailing the characteristics of the cell nuclei visible in the image [31]. We evaluated and compared the performance of five classification algorithms: Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machines, and Random Forest, based on four performance metrics: Sensitivity, Specificity, Precision, and Recall. The main aim is to determine which algorithm would yield the highest accuracy when applied to our dataset.

Data set Information

- Attribute information: ID number, Diagnosis (M = malignant, B = benign).
- Ten real-valued features are computed for each cell nucleus:
- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter∧2 / area - 1.0)
- concavity (severity of concave portions of the contour)

- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension (“coastline approximation” - 1)

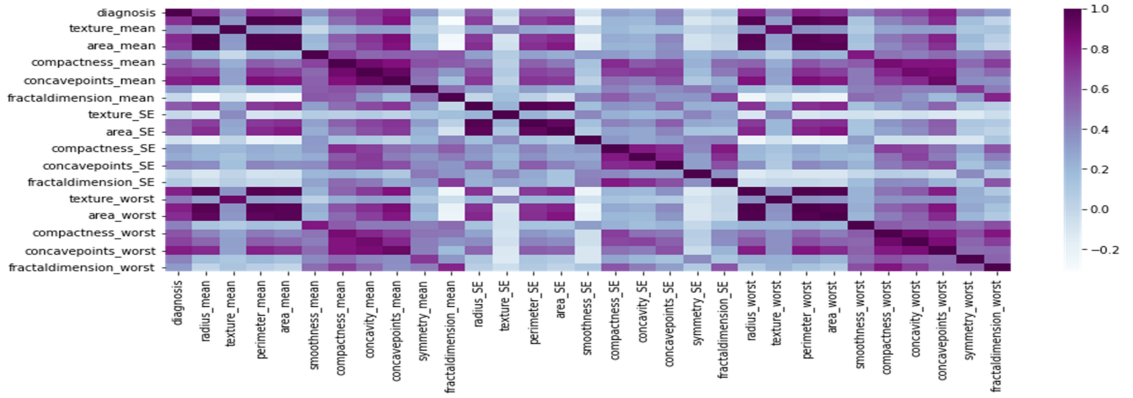
To ensure data quality, the dataset was cleansed to eliminate redundant data and missing values. Unnecessary columns were removed, and Pearson’s coefficient was used to determine the pair-wise correlation among all 31 columns, excluding the ID number column. As expected, the diagonal values on the heatmap in FIGURE 3 (i), were all 1, indicating the correlation of a variable with itself.

Following data cleansing, the next step is data normalization, as different features may have different scales, which is not ideal for machine learning algorithms. It is crucial to bring each feature to an optimal range for accurate analysis. One commonly used approach is Min-Max normalization, which transforms every feature’s value to a decimal between 0 and 1 [32], with the minimum value being 0 and the maximum being 1. However, it is important to note that this approach does not handle outliers well, even though it guarantees that all features will be of the same scale. Additionally, the categorical outcome of disease diagnosis, represented by “M” and “B”, was converted to numerical values of 1 and 0, respectively, as illustrated in FIGURE 3 (ii).

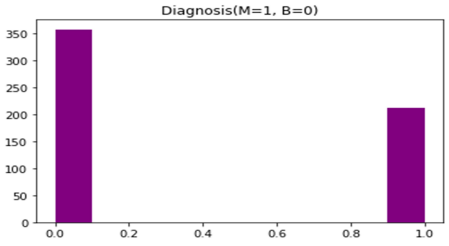
Based on FIGURE 3 (iii), it can be seen that the means of the radius, perimeter, and area exhibit linear patterns, indicating that these features are highly correlated and therefore multicollinear. Multicollinearity occurs when there are high inter-correlations between independent variables in a multiple regression model. This phenomenon may lead to misleading results when attempting to analyze how each independent variable can predict the dependent variable in a statistical model [33].

FIGURE 3 (iii), illustrates the effect of multicollinearity, as the radius mean column has a higher correlation of 1.0 and 0.99 with the perimeter mean and area mean columns, respectively. When performing correlation-based feature selection, it is necessary to keep only one feature from the remaining three. Similarly, the area mean column has a higher correlation of 0.96 with each of the radius worst, perimeter worst, and area worst columns.

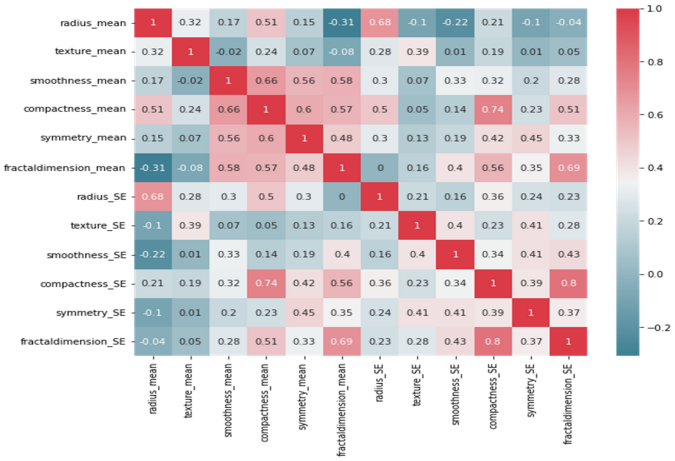
After dropping the required number of columns, we visualized the correlation matrix containing 12 features, including radius mean, texture mean, smoothness mean, compactness mean, symmetry mean, fractal dimension mean, radius SE, texture SE, smoothness SE, compactness SE, symmetry SE, and fractal dimension SE. The correlation matrices are shown in FIGURE 3 (iv), and FIGURE 3 (v), respectively. The data visualization was performed using popular Python libraries such as matplotlib, seaborn, NumPy, and pandas. To better understand the data, we utilized various histogram plots, including heatmaps and violin plots. To overcome the issue of multicollinearity, one effective technique is to use principal component regression (PCR) [34], a regression analysis method based on principal component analysis (PCA).



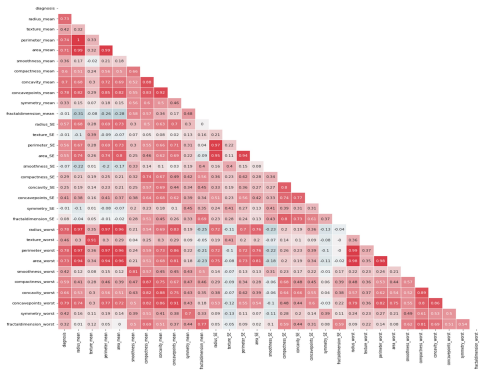
(i) Pair wise Correlation among 31 columns



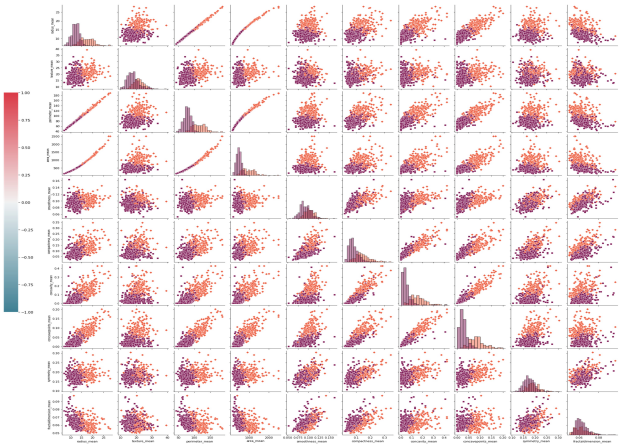
(ii) Histogram showing the disease diagnosis (M =212, B=357)



(iii) Correlation Matrix with 12 extracted features



(iv) Correlation Matrix obtained by masking the upper triangle



(v) Matrix extracting the "Mean" columns

Figure 3: Correlation Matrix with 12 extracted features

3.1.2 Experimental evaluation and results

PCA is obtained by preprocessing and scaling the model using the standard scaler from the sk-learn library in Python, which extracts the top 12 features out of the 31 features available in the dataset. By selecting 12 features, we retain 97% of the total variance of the data. To evaluate the effectiveness of the five classification algorithms, namely Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest (RF), we apply n-fold cross-validation with n=10, where we divide the dataset into n-folds and train on (n-1) folds. The performance of each algorithm is evaluated based on parameters such as i. Correctly classified instances, ii. Incorrectly classified instances, iii. Accuracy, as shown in TABLE 3. Furthermore, Figure 4, presents a graph comparing the performance of each classifier using PCA feature selection with the top 12 features. To evaluate the performance of each classification algorithm, we measure four metrics: i. Precision (Positive Predicted Value), ii. Recall (Sensitivity, True Positive Rate), iii. F1-score (Measures both precision and recall), and iv. Accuracy. The performance results are shown in TABLE 3.

Table 3: Effectiveness of the classification algorithms using PCA feature selection with top 12 features.

Classifiers (Using PCA)	LR	NB	KNN	SVM	RF
Evaluation Parameters	(k = 4)				
Correct instances	551	544	551	549	555
Incorrect instances	18	25	18	20	14
Accuracy (%)	96.8	95.7	96.8	96.4	97.2

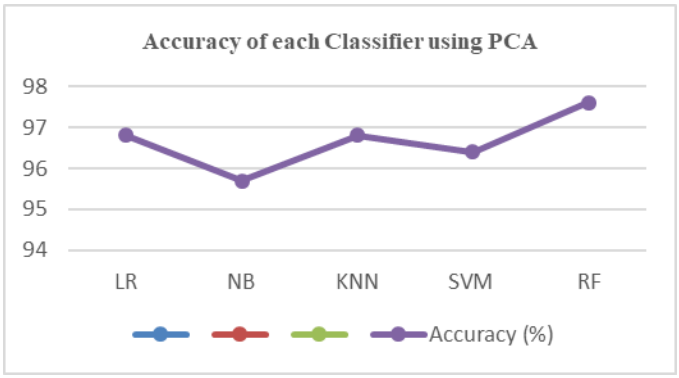


Figure 4: Graph comparing accuracy of each classifier using PCA feature selection with top 12 features.

3.1.3 Performance metrics to assess quality of each classification algorithm

The confusion matrix of each classifier allows us to identify how efficiently the model performs in identifying the WDBC disease. The four parameters of the confusion matrix include i. True Positive (TP), ii. True Negative (TN), iii. False Positive (FP), and iv. False Negative (FN). Based

on these four parameters, we evaluate the four-performance metrics, i.e., i. Precision, ii. Recall, iii. F1-score, and iv. Accuracy.

1. Precision (Positive predictive value), the ratio of all positive instances which are correctly classified, is defined as:
$$Precision = TP/(TP + FP)$$
2. Recall (Sensitivity, True Positive Rate) is the ratio of true positives defined as:
$$Recall = TP/(TP + FN)$$
3. F1-score is a combined measure that trades off both precision and recall. Both precision and recall must be good to achieve a high F1-score and is defined as:
$$\boldsymbol{F}1 - \boldsymbol{score} = \frac{2 \times \boldsymbol{precision} \times \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}}$$
4. Accuracy, the ratio of correctly classified instances is defined by using the formula:
$$Accuracy = \frac{TP + TN}{All\ data},\ where\ All\ data = TP + FP + TN + FN$$

It can be seen from TABLES 3 and 4, and FIGURE 4, and FIGURE 5, that the Random Forest (RF) classifier achieves an accuracy of 97.2% and outperforms other classifiers in terms of effectiveness and efficiency. This result is obtained using principal component analysis (PCA) feature selection with the top 12 features. It is worth noting that the RF classifier has a low correlation compared to other classifiers, making it an effective choice. Models with low correlation generally provide more accurate predictions. Therefore, the RF classifier outperforms other classifiers in terms of precision, accuracy, recall, and F1 score in classifying breast carcinoma on the WDBC dataset.

Table 4: Efficiency of the different performance metrics

Classifiers	Precision	Recall	F1-score	Support	Diagnosis
LR	0.94	0.97	0.96	67	Benign
	0.98	0.97	0.97	121	Malignant
NB	0.94	0.94	0.94	72	Benign
	0.97	0.97	0.97	116	Malignant
KNN	0.97	0.94	0.96	72	Benign
	0.97	0.98	0.97	116	Malignant
SVM	0.94	0.97	0.95	63	Benign
	0.98	0.96	0.97	108	Malignant
RF	0.93	0.96	0.95	57	Benign
	0.98	0.96	0.97	114	Malignant

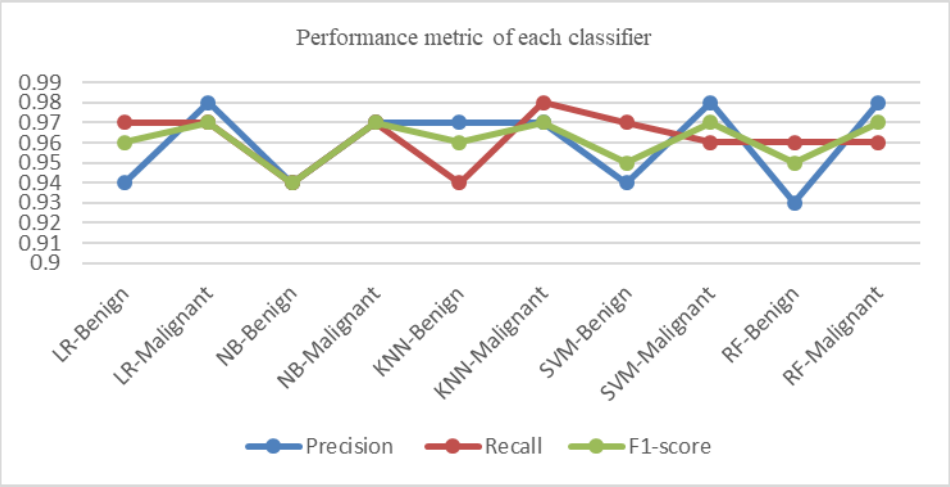


Figure 5: Graph comparing the performance metric of each classifier.

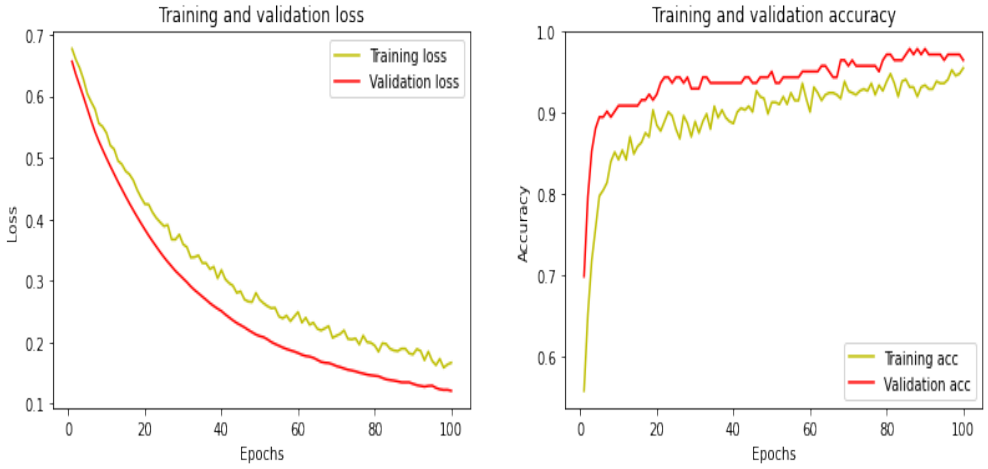


Figure 6: Graph comparing the training and validation loss and training and validation accuracy of RF classifier.

3.2 Particle Swarm Optimization

PSO is an evolutionary algorithm that was proposed by Kennedy et al. in 1995. This optimization technique is based on parallel stochastic search and uses an analogy with insect or bird swarms [35]. Traditional feature selection methods often encounter issues related to high computational costs and the risk of becoming stuck in local optima. To overcome these issues, global search techniques, such as evolutionary computation methods like PSO and genetic algorithms (GA), are used for feature selection [36–38]. These algorithms are considered methods for design, optimization, and problem solving that mimic the process of natural evolution to some degree. Compared to other methods,

PSO is computationally less expensive and has the potential to converge more quickly [39], offering promising solutions to feature selection problems.

PSO is an optimization technique that aims to improve candidate solutions iteratively based on a given quality measure. The algorithm uses a population of randomly generated particles X on an n-dimensional search space and an objective function F defined on X, which maps these n-dimensional particles to a real number R. The objective of PSO is to locate a position in the search space where the objective function can be maximized. The PSO algorithm [40], works by:

i Defining a swarm of randomly placed particles $X = \{x1, x2, \ldots, xn\}$.

ii Allowing the particles to move in the search space based on their own experience and that of other particles.

iii Collaboratively converging to a near-optimal location.

Each particle in PSO is defined by its velocity and position, and each particle corresponds to a solution that is randomly generated. The goal of PSO is to search for a near-optimal solution in the search space by iteratively changing the particles' velocity and position, finding, and updating its personal best position (p-Best) and global best position (g-Best) if needed [41]. Essentially, each particle moves towards a combination of its own remembered best position and the remembered global best position at each time step. The particles in PSO fly through the problem space by following the current optimum particles. The algorithm aims to find the best particle, i.e., the one with the highest fitness value, and then update the swarm's position to converge to a near-optimal location. PSO is a highly efficient optimization algorithm that has been widely used in many different applications, such as data mining, image processing, and financial forecasting, among others. Appendix A describes the PSO algorithm.

3.2.1 Experimental evaluation and results

We conducted an evaluation of the effectiveness of five classification algorithms using PSO feature selection on the WDBC dataset. PSO was used to capture the top 15 features out of the 32 available in the dataset, which were then tested on each classifier to assess their effectiveness based on four parameters: (i) correctly classified instances, (ii) incorrectly classified instances, and (iii) accuracy of each classifier.

In addition to assessing effectiveness, we also evaluated the efficiency of different performance metrics of each of the five classification algorithms. This evaluation was based on four metrics: (i) precision (positive predicted value), (ii) recall (sensitivity, true positive rate), (iii) F1-score (which measures both precision and recall), and (iv) support. TABLE 5, displays the results of this evaluation for each of the performance metrics.

Based on the results presented in TABLE 5, and FIGURE 7, we can observe that the Random Forest (RF) classifier achieves an impressive accuracy of 95.74% and outperforms the other classifiers in terms of effectiveness and efficiency, as also illustrated in TABLE 6, and FIGURE 8, using the particle swarm optimization (PSO) feature selection with the top 15 features. It is noteworthy that the

Table 5: Effectiveness of the Classification Algorithms using PSO Feature Selection with top 15 features

Classifiers (Using PCA)	LR	NB	KNN	SVM	RF
Evaluation Parameters	(k = 4)				
Correct instances	534	529	524	524	545
Incorrect instances	35	40	45	45	24
Accuracy (%)	93.85	93.08	92.10	92.10	95.74

RF classifier shows low correlation when compared to other classifiers, making it a more effective model for accurate predictions [42]. Therefore, the RF classifier outperforms other classifiers with respect to precision, accuracy, recall, and F1 score in the classification of breast carcinoma.

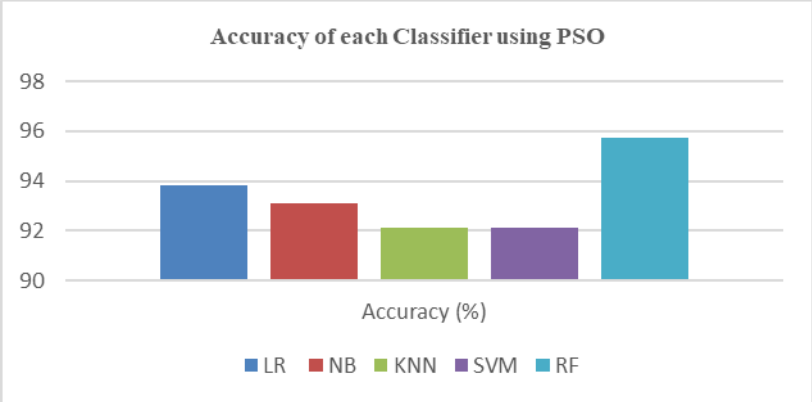


Figure 7: Graph comparing each classifier using PSO Feature Selection with top 15 features.

Table 6: Efficiency of the different performance metrics

Classifiers	Precision	Recall	F1-score	Support	Diagnosis
LR	0.95	0.90	0.93	42	Benign
	0.95	0.97	0.96	72	Malignant
NB	0.93	0.86	0.89	72	Benign
	0.92	0.96	0.94	116	Malignant
KNN	0.97	0.81	0.88	42	Benign
	0.90	0.99	0.94	72	Malignant
SVM	1.00	0.76	0.86	42	Benign
	0.88	1.00	0.94	72	Malignant
RF	0.98	0.89	0.93	72	Benign
	0.93	0.99	0.96	116	Malignant

The training, validation, and testing were performed using a 70-10- 20 percent split for both feature selections PCA and PSO, employing the random forest (RF) classifier. FIGURE 6, and FIGURE 9, illustrate the loss and accuracy curves obtained for both training and validation. Both the training and validation loss began with a higher value and continued to decrease as the training progressed. Notably, the validation loss was lower than the training loss, even without using a dropout, indicating

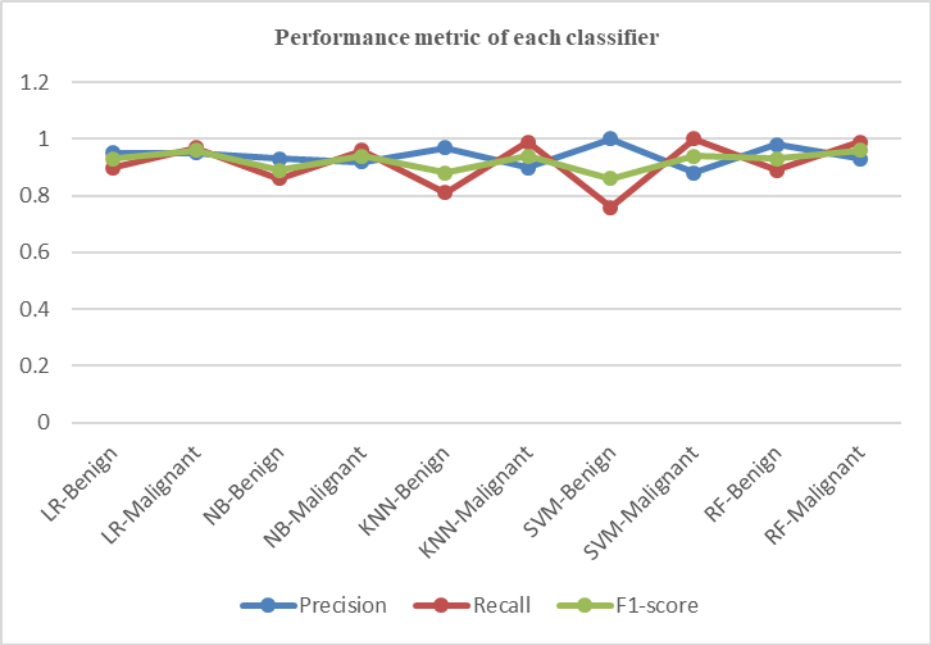


Figure 8: Graph comparing the performance metric of each classifier.

that the model became more robust during testing than it was during training, resulting in higher testing accuracies. As seen in the figures, both the training and validation accuracies increased with an increase in the number of epochs, reached a plateau, and then saturated. Therefore, the Random Forest model overcame the issues of both underfitting and overfitting, demonstrating its efficiency and effectiveness in identifying the breast cancer type.

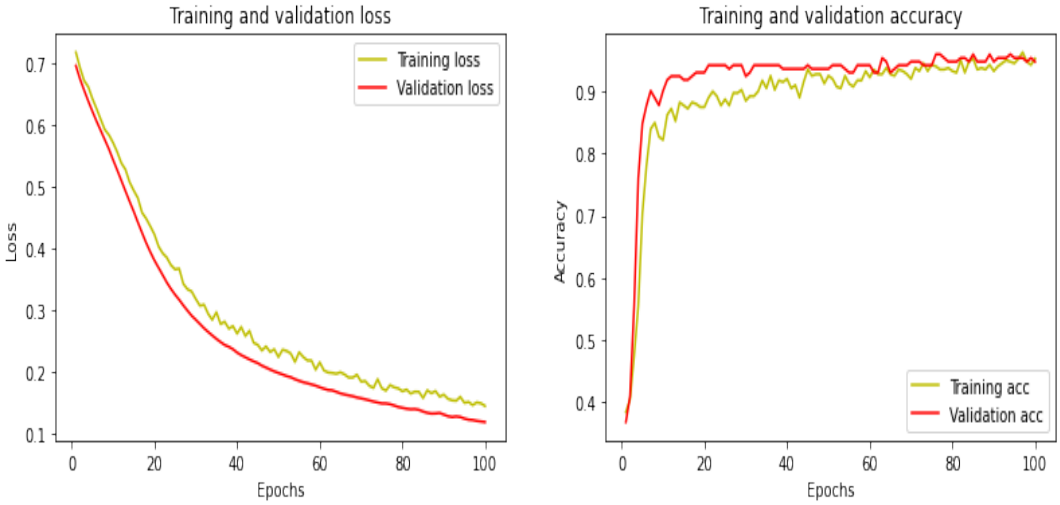
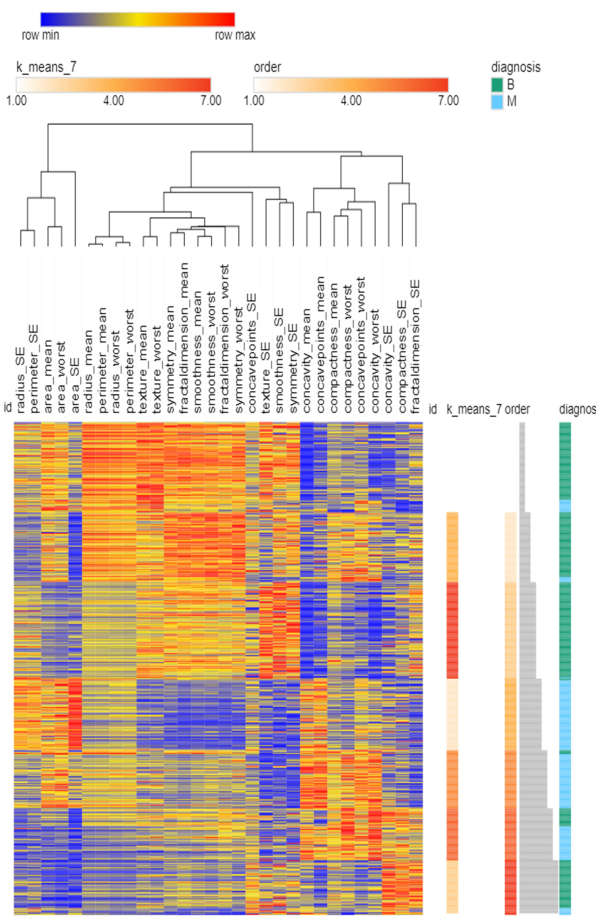


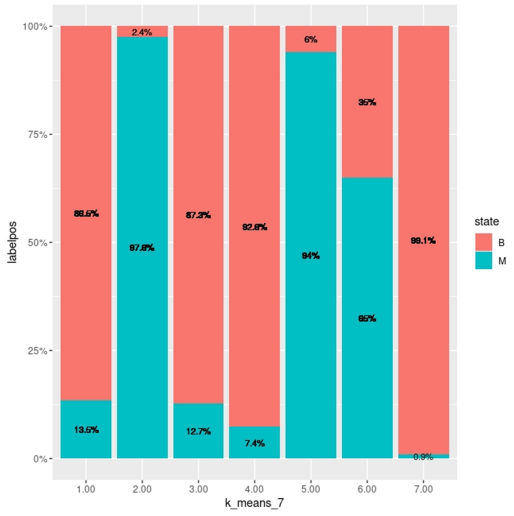
Figure 9: Graph comparing the training and validation loss and training and validation accuracy of RF classifier.

3.2.2 Interpretability of clinical data using k-means

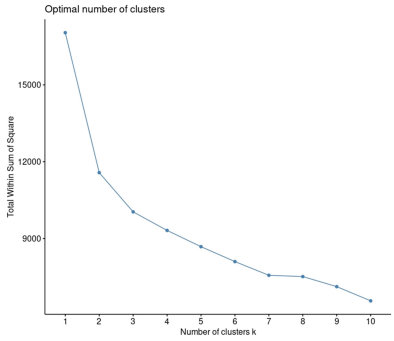
The goal of the analysis is to cluster the patients sample IDs based on the available features and cluster these into "benign" and "malignant" groups which can be seen as a binary classification problem. Two clustering analysis approaches – Hierarchical and Non-Hierarchical clustering (ex. K-means) were used. FIGURE 10 (i), uses a K-Means clustering created using Morpheus [43], a data visualization and analysis tool to visualize the dataset as a heat map and then perform some exploratory analysis - in our case, K-Means.



(i) k-Means clustering showing the percentage of cells



(ii) Barplot representing the % of Malignant vs Benign cells in each of the k clusters



(iii) Optimal number of clusters

Figure 10: K-Means showing the percentage of cells in each of the K clusters (K=7)

The different feature patterns are demonstrated by the different malignant or benign samples as seen in FIGURE 10 (i). The number of clusters is seven from k=1 to 7. One can even go higher with the different values of k to obtain better metrics and distributions. For example, for k=1, 2, 3, 4, 5, 6;

the diagnosis looks uniform and almost entirely benign and the samples that are benign with these structures are very different than the other representations of k.

FIGURE 10 (ii), is a bar plot representing the percentage of Benign and Malignant states in each of the k clusters ranging from 1 to 7. For k = 1, 3, 4, 7, we see a higher percentage of Benign cells i.e., 86.5%, 87.3%, 82.6% and 98.1% respectively. On the other hand, for k = 2, 5, 6, we see a higher percentage of Malignant cells i.e., 97.8%, 94%, and 86%. This bar plot has been extracted from a JSON file after saving a session from FIGURE 10 (i). This way one can view this visualization as a guide to obtain the informative categorization of the patient diagnosis outcomes obtaining more reasonable clusters from a practical standpoint than the other available statistical techniques [44]. The optimal number of clusters obtained is seen in FIGURE 10 (iii), the total within sum of squares shows the amount of variation on the dependent variable.

3.2.3 Discussion and Key observations

Sections 3.1 and 3.2 evaluate various classification algorithms, their effectiveness and efficiency in diagnosing the breast cancer type, using PCA and PSO feature selection techniques. Feature selection is a crucial factor in machine learning, as it significantly impacts the accuracy and performance characteristics of different classification algorithms. It is seen that Random Forrest Classifier outperforms the other classification algorithms and achieves an overall accuracy of 97.2% and 95.7% respectively with PCA and PSO feature selection techniques. We observe that the other four classification algorithms also had accuracies over 90%. The main aim in evaluating the models is to maximize the classification performance with a minimal set of important features. In comparison using PCA feature selection on RF classifier gives a better classification accuracy of 97.2% with a reduced set of 12 features while on the other hand, using PSO feature selection on RF classifier gives a classification accuracy of 95.7% with a 15-feature set, out of the 32 features shown in TABLE 7. Therefore, it can be concluded that the PCA feature selection is more efficient than PSO. These results suggest that, as part of future analysis, one can test these proposed techniques with much larger datasets and with a much larger number of features.

Table 7: Comparison of RF classifier accuracies with PCA and PSO

Classifiers	Feature Selection	No. of Reduced Features	Approach	Accuracies
RF	PCA	12	RF + PCA	97.2%
	PSO	15	RF + PSO	95.7%

The performance metrics of the Random Forest classifier on the WDBC data set are as follows: Benign (Precision-93%, Recall-96%, F1-score-95%) and Malignant (Precision-98%, Recall-96%, F1-score-97%) for PCA and (Precision-98%, Recall-89%, F1-score-93%) and (Precision-93%, Recall-99%, F1-score-96%) for PSO. Section 3 also throws light on the interpretability of WDBC clinical data and performs exploratory analysis using k-Means to see the percentage of benign and malignant cells in each of the k clusters.

4. FEATURE EXTRACTION FROM WHOLE SLIDE IMAGES

In recent years, efforts in the digitization of WSI's have gained immense prominence in the field of medical image analysis and have opened doors for studying and developing various deep learning models for analysis, prediction, and treatment of various types of carcinomas. Tissue samples on a whole slide are diagnosed by using the staining elements namely H&E (hematoxylin and eosin). The biological cell structures are visualized under a microscope with the light illuminating from below the slide of stained tissues or using images of high resolution namely digitized WSI. When there is no stain, a bright white light appears and the whole light is passed through the slide. The amount of light absorbed by a stain indicates that the stain has adhered to a substance and had absorbed some of the light. This is an old technique to predict cancer cells from H&E-stained tissues and has many limitations. This includes cancer cells that may have multiple appearances and cells may exhibit similar hyperchromatic features thereby making the prediction difficult. To address these concerns, deep learning models help in better visualization of image patterns so that even the tiniest dot isn't missed in the identification. There have been various deep learning methods used so far to predict tumor type. This prediction is either a binary classification or multi-class classification [44], problem. One can use pretrained networks such as VGG-16, Inception-V3, Inception-V4, U-NET, Mask RC-NN, to name a few. Deep learning algorithms make use of a sequence of tasks that include preprocessing of images, segmentation, feature extraction through convolutional neural networks (CNNs), and finally classification (either benign or malignant). These algorithms help in assisting the pathologist to detect the cancer subtypes and gene mutations. Additionally, with the rapid advancements in AI tools there is an increasing demand for predictive assays that help in the additional treatment of patients during surgery [45].

4.1 The WSI Data Set

Histopathological Breast WSI's have been derived from The Cancer Genome Atlas (TCGA) portal. The project name is Breast Invasive Carcinoma to detect the disease type Ductal and Lobular Neoplasms. The samples have been derived through SOB (surgical open biopsy). These WSI's have been taken from the vendor Aperio on an objective power of 40X (the slide image is captured at 40X resolution), obtained using the Open Slide library. WSIs resembles a pyramid structure with different levels of resolution. Considering an example of one such image, the slide dimensions at the native resolution are 44743×50293 pixels but a typical whole slide image may contain 100,000×100,000 pixels. One can get the slide dimensions at each level by obtaining the number of levels in the WSI and acquiring the dimensions of various levels followed by down sampling each level by a certain amount. The WSI are converted to RGB channels and then, at the implementation level, to a NumPy array for further processing. Since WSIs are large images, it is important to divide them into several tiles for deep learning training or other forms of processing.

FIGURE 11, displays the tiles extracted from a given WSI. A common method of visualizing the sample is to stain them using H&E - the preferred ways to stain images by pathologist. Depending on how round or dense the nuclei are, it enables the pathologists to interpret a given tissue. The H component selectively stains the nucleic acids a blue-purple and the E component stains the proteins with a bright pink color. Using the algorithm described in [46], as an example, we extract a small patch/tile from a given WSI and obtain the normalized image, the H image and the E image as shown

in FIGURE 12. This algorithm converts a given RGB image to an optical density (OD) and removes data with the intensity of OD < β (threshold, β = 0.15). Next step in the process is to calculate the singular value decomposition (SVD) on the OD image and create a plane from directions obtained from SVD that correspond to the two largest singular values. The data is then projected onto the plane and normalizes to unit length. In context to the SVD direction, the algorithm then calculates the angle of each point to obtain the robust extremes of the angle (α^{th} and (100- α)th percentiles). Finally, it converts these extremes back to the optical density (OD) space to obtain the optical stain vectors.

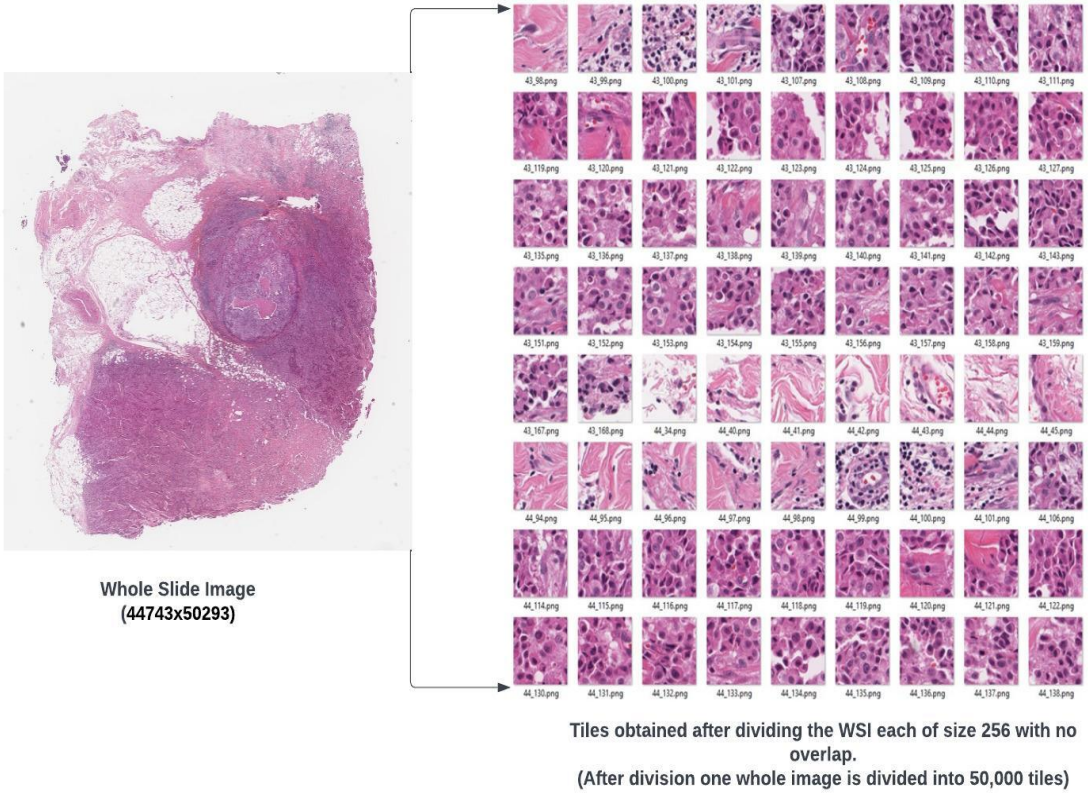


Figure 11: Tiles extracted from a given WSI using the Deep Zoom Generator from Open Slide

FIGURE 12, illustrates the process of staining the slides using H&E but this technique is prone to intra and inter observer variability and suffers from low throughput [47]. Because of this reason, a growing interest in digital pathology has acquired much attention wherein the digitized WSI's are captured from glass slides using a scanning device. This procedure therefore allows efficient processing and analysis of tissue specimens [48]. WSI includes tens of thousands of nuclei of different types (cell types) that can be analyzed to predict the disease outcome. Based on the nuclear features, one can predict the survival rate, tumor grade, disease type. It is important to note that good quality tissue segmentation involves efficient and accurate prediction. Nuclei segmentation is the initial step for further downstream analysis to assess and visualize the tissue components that contribute to the disease [49]. It is worth remembering here that nuclei display a high level

of heterogeneity in terms of its shape, size, and the chromatin patterns between different cell and disease types.

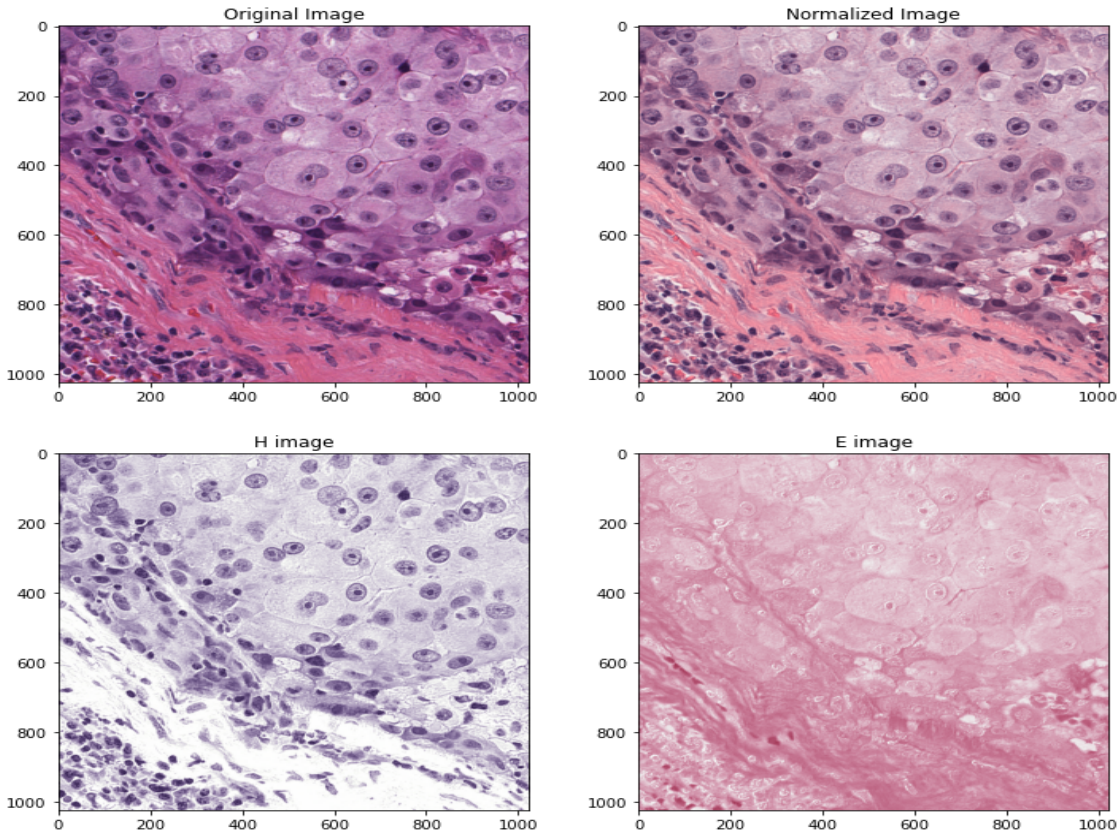


Figure 12: Normalized image, the H image and the E image for a given tile/patch.

4.2 Nuclei Segmentation in WSIs

For many biological applications, it is important to detect and segment cells and nuclei. The main idea of segmentation is to classify cells at a pixel level. There are two types of segmentation techniques: semantic segmentation and instance segmentation. Recent approaches made use of per-pixel cell segmentation that include grouping of pixels and the use of bounding box with refinement of shapes. This technique may suffer from segmentation errors in situations where there are crowded cells. On the other hand, instance segmentation is the process of assigning a cell instance identity to each pixel in the image [50]. In the bottom-up approach, each pixel is first classified into certain semantic classes i.e., either cell or background and then pixels are grouped to individual instances if they fall under the same class. This approach makes use of learned classifiers, for example random forests, or various types of neural networks.

Although this approach achieves good results, it has issues for images in situations where crowded cell nuclei are present. To address these concerns, a top-down approach is suggested wherein the individual cell instances are first localized with some shape and then the shape is refined using

object detection methods that predict and classify the pixels within each box, i.e., axis aligned bounding box (e.g., mask-RCNN). Such methods prevent detecting the same object multiple times in situations where boxes with lower confidence are suppressed by boxes having higher confidence using an NMS (non-maximum suppression) step. If the objects are poorly represented by their axis-aligned bounding boxes, NMS is problematic. To address this, the approach in [51], makes use of rotated bounding boxes but it becomes necessary to accurately refine the box shape to distinguish the objects, for example, the cell nuclei. To address all the concerns, especially to get rid of the crowded nuclei, that cause merging bordering cell instances, the authors proposed a StarDist model. StarDist is a cell detection approach that is successful in predicting the shape representation that is flexible and whose accuracy can compete with that of the instance segmentation methods. The star-convex polygons used in [51], approximate the typically roundish shapes of cell nuclei in microscopy images. The architecture used by the StarDist model [52], makes use of a light-weight neural network based on U-Net [53], which is easy to train with state-of-art approaches. Appendix B reviews the U-net architecture.

4.3 Cell Detection Using Stardist

StarDist is a pretrained model that makes use of a light-weight neural network based on U-Net that is easier to train and exhibits a good performance compared to other state of art methods. StarDist uses object detection similar to that presented in [54], that can predict shapes for each interested object. No axis-aligned bounding box is used because of the limitations discussed in [55].

4.3.1 The working of a starDist model

The purpose of the model is to predict a star-convex polygon such that for each (i, j) image pixel a set of n radial distances from the center to the object boundary are computed at equidistant angles. The model separately predicts the object probability $d_{i,j}$ for each pixel (i, j). This way all the polygons are predicted with their respective object probabilities by performing a non-maximum (NMS) suppression. Each polygon represents an individual instance of the object. In conclusion, both the object probabilities that include the Euclidean distance to the nearest background pixel that supports polygons associated to the nearest cell center, and the computation of the star-convex polygon distances $r^k_{i,j}$ to the object boundary by following the radial directions k until it encounters a different object identity. The GPU implementation is carried out on GoogleColab to compute the distances on demand.

FIGURE 13, illustrates a general process for 2D images, where the training data consists of pairs of raw images input with fully annotated labels meaning that each pixel is labeled either with a unique object id, or a 0 for the background. This model is then trained to densely predict the radial distances to the object boundaries with object probabilities d to output a set of candidate polygons for a given input image.

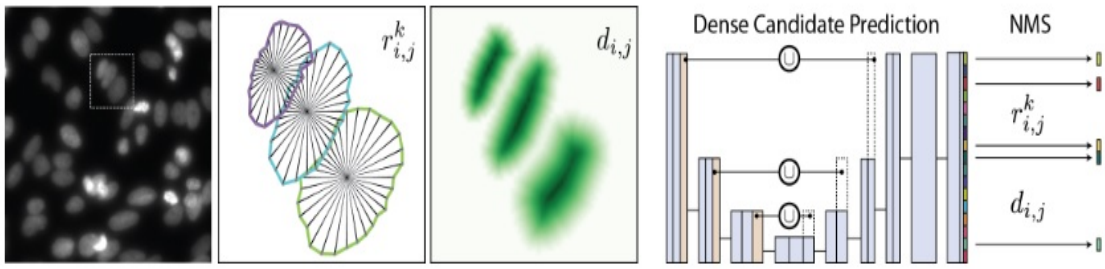


Figure 13: Process carried using 2D images [56]

4.3.2 Experimental results

First, the given histopathological image is rendered at different resolutions as seen in FIGURE 14. Using a pretrained model, StarDist uses a normalizer to predict the instances from a WSI, that contain hundreds and thousands of nuclei as shown in FIGURE 15.

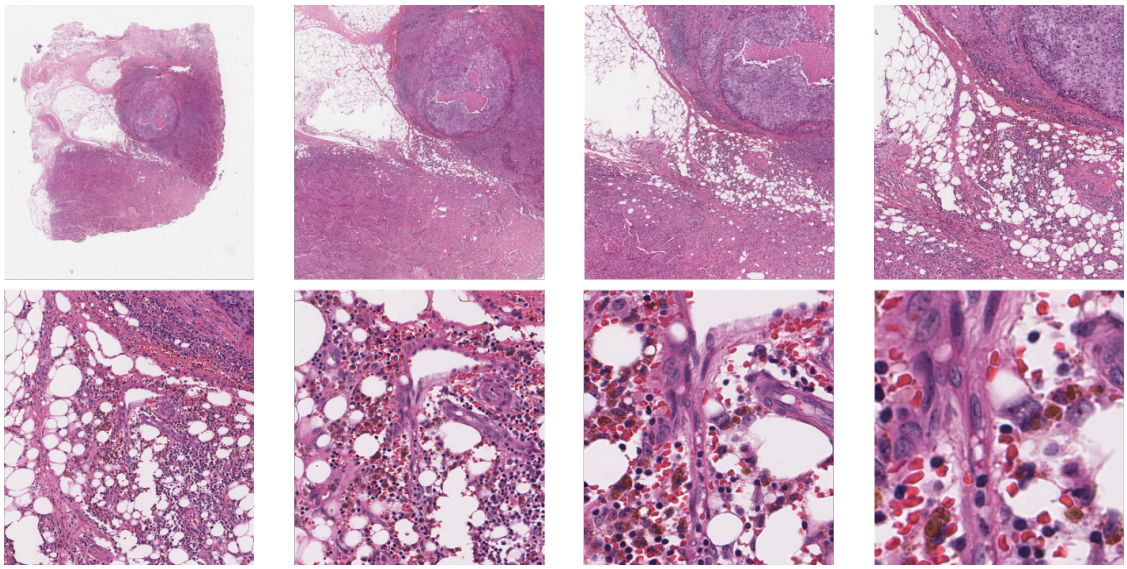


Figure 14: WSI displayed on different resolutions.

The model is trained on segmented input images each of size 256×256. Instances are randomly predicted, and the radial distances are computed by visualizing each segmented nuclei as a polygon. One such scenario of the instances predicted for a given image tile is seen on the right of FIGURE 16. The radial distances and the object boundaries are plotted as shown in the left side of FIGURE 16.

The image is quantified to extract information regarding the size distribution of each segmented nuclei which involves extracting the relevant features. To extract these metrics, a *regionprops_table* is used as part of the sciKit image library, which is then converted into a Pandas data frame. This way we provide an original image, the labelled image, to extract the intensity information for each of the objects as shown in TABLE 8.

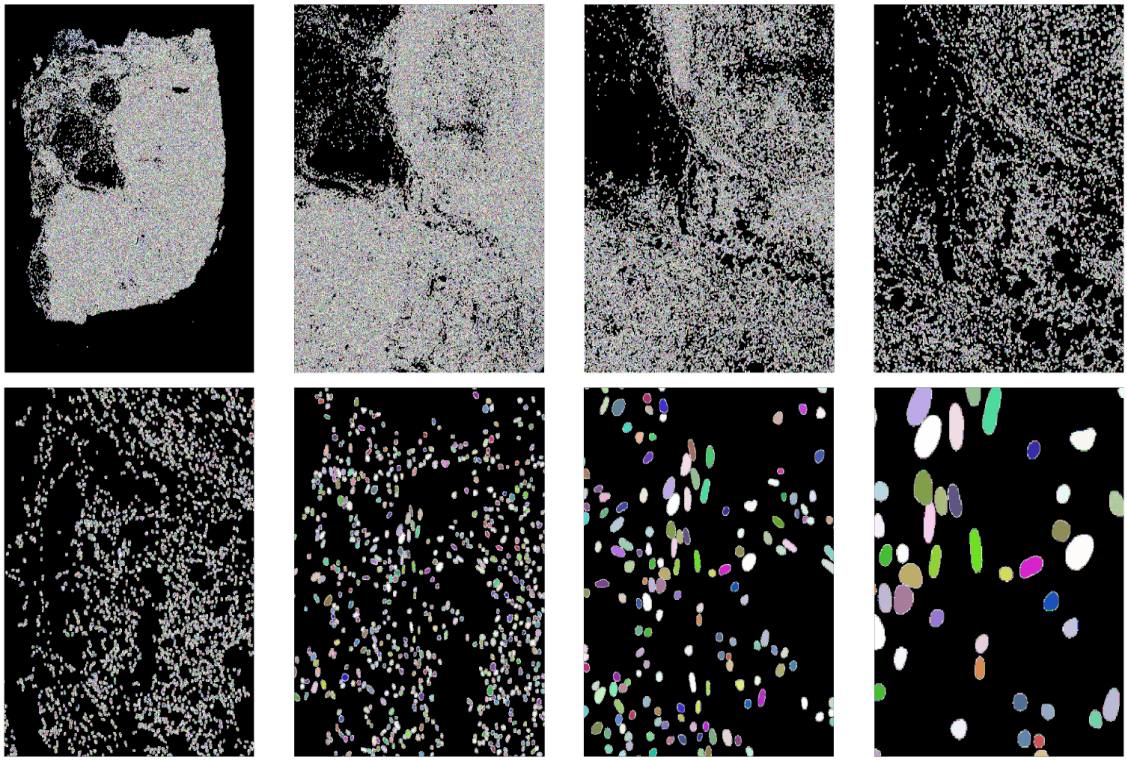


Figure 15: Displaying the nuclei using instance segmentation for a given WSI at different resolutions

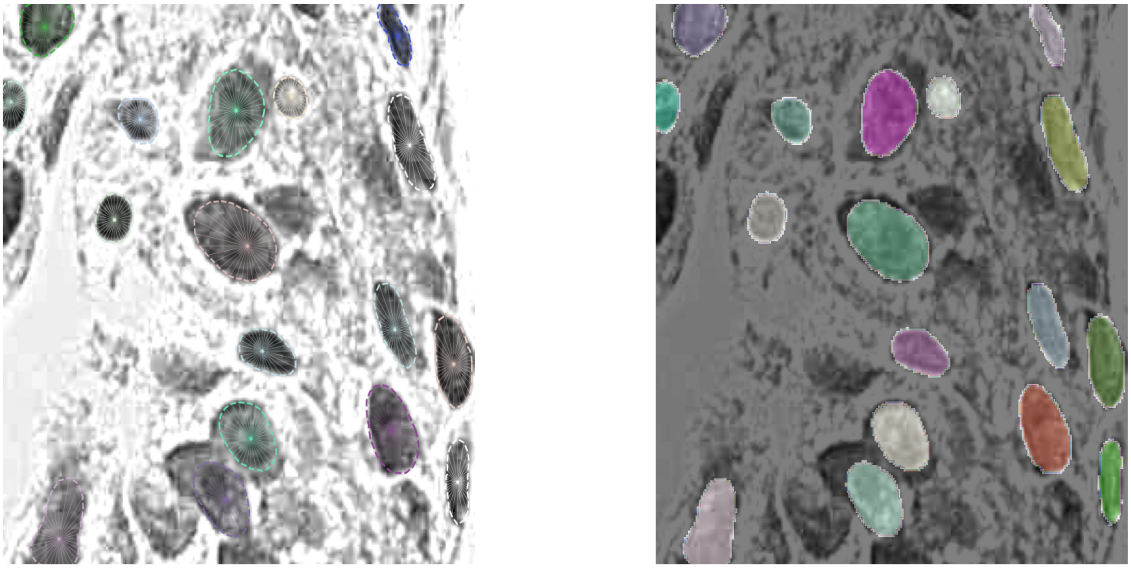


Figure 16: Displaying the nuclei using instance segmentation for a given WSI at different resolutions

Table 8: Size distribution of the Segmented Nuclei

	label	area	equivalent diameter	mean_ intensity-0	mean_ intensity-1	mean_ intensity-2	solidity
0	1	416	23.014510	113.033654	90.451923	130.322115	0.962963
1	2	302	19.609139	89.069536	72.986755	111.450331	0.943750
2	3	502	25.281738	129.153386	96.880478	134.442231	0.965385
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
227896	227897	484	24.824342	76.514463	61.280992	71.159091	0.954635
227897	227898	480	24.721549	106.600000	112.468750	105.008333	0.958084

5. CONCLUSIONS AND FUTURE DIRECTIONS

This study focused on feature selection techniques on both clinical data and whole slide images. As part of the clinical data, we used two feature selection techniques such as PCA and PSO to select the top features and evaluated these on 5 different classifiers. The results show that the random forest (RF) classifier with top 12 PCA features achieved the highest accuracy of 97.2% as compared to RF classifier with top 15 features that achieved an accuracy of 95.7%. The second half of the paper discussed the feature selection techniques for WSI's from The Cancer Genome Atlas (TCGA). We discussed StarDist that uses a deep convolutional neural network (U-Net) as a backbone model that is trained on WSI's to extract relevant features. This paper discussed feature extraction, nuclei-based instance segmentation, H&E-stained image extraction, and quantifying intensity information for a given WSI. Current image analysis is becoming more useful for prediction of the underlying genomic data from single cell genomics and of the cancer-causing mutations. Going forward, the focus of this research is how to predict tumor sensitivity to certain treatments based on tumor histology.

References

[1] https://www.cdc.gov/cancer/breast/statistics/

[2] https://www.who.int/news-room/fact-sheets/detail/breast-cancer/

[3] https://builtin.com/data-science/step-step-explanation-principal-component-analysis/

[4] Noble WS. What Is a Support Vector Machine? Nat Biotechnol. 2006;24:1565-1567.

[5] Rish I. An Empirical Study of the Naive Bayes Classifier.IJCAI Work Empirical Methods Artif Intell. 2001;3:41-46.

[6] Larose DT, Larose CD. k-Nearest Neighbor Algorithm. Discovering knowledge in data: An Introduction to Data Mining. Hoboken, NJ: John Wiley & Sons, Inc. 2014.

[7] Subasi A. Practical Machine Learning for Data Analysis Using Python. Academic Press. 2020;465-511.

[8] Masetic Z, Subasi A. Congestive Heart Failure Detection Using Random Forest Classifier. Comput Methods Programs Biomed. 2016;130:54-64.

[9] Xue B, Qin AK, Zhang M. An Archive-Based Particle Swarm Optimization for Feature Selection in Classification. In: Proceedings of the 2014 Congress on Evolutionary Computation, Beijing: IEEE Publications. 2014:3119-26.

[10] Mohemmed AW, Zhang M, Johnston M. Particle Swarm Optimization Based Adaboost for Face Detection. In: Proceedings of the 2009 Congress on Evolutionary Computation. Trondheim: IEEE Publications. 2009:2494-2501.

[11] Xue B, Zhang M, Browne WN. New Fitness Functions in Binary Particle Swarm Optimization for Feature Selection. In: Proceedings of the 2012 Congress on Evolutionary Computation. Brisbane, QLD: IEEE Publications. 2012:1-8.

[12] Yeoman TB, Xue B, Zhang M. Particle Swarm Optimization for Feature Selection: A Size Controlled Approach. Proceedings of the 13th Australasian Data Mining Conferences on Research and Practice in Information Technology. 2015;151-159.

[13] Zhou Y, Onder OF, Dou Q, Tsougenis E, Chen H, et al. CIA-net: robust nuclei instance segmentation with contour-aware information aggregation. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019:682-693.

[14] Raju A, Yao J, Haq MMH, Jonnagaddala J, Huang J. Graph Attention Multiinstance Learning for Accurate Colorectal Cancer Staging. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2020:529-539.

[15] Saillard C, Schmauch B, Laifa O, Moarii M, Toldo S, et al. Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides. Hepatology. 2020;72:2000-2013.

[16] Asri H, Mousannif H, Moatassime HA, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Comput Sci. 2016;83:1064-1069.

[17] Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge based Analysis of various statistical tools in detecting breast cancer. Comput Sci Inf Syst. 2011;2:37-45.

[18] Bhardwaj A, Bhardwaj H, Sakalle A, Uddin Z, Sakalle M, et al. Tree-based and machine learning algorithm analysis for breast cancer classification. Comput Intell Neurosci. 2022.

[19] Liu Y, Wang CW, Zhang L. Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data. 3rd International Conference on Bioinformatics and Biomedical Engineering. IEEE Publications. 2009:1-4.

[20] Agarap AFM. On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset. In: Proceedings of the 2nd international conference on machine learning and soft computing. 2018:5-9.

[21] Sankari L, et.al. Predicting Breast Cancer Using Novel Approach in Data Analytics. International Journal of Engineering Research And. 2017;6.

[22] Das K, Conjeti S, Chatterjee J, Sheet D. Detection of breast cancer from whole slide histopathological images using deep multiple instance CNN. IEEE Access. 2020;8:213502-213511.

[23] Lu W, Graham S, Bilal M, Rajpoot N, Minhas F. Capturing cellular topology in multi-gigapixel pathology images ieee/Cvf conference on computer vision and pattern recognition workshops. 2020:260-261.

[24] Hall MA, Smith LA. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper in proceedings of the twelfth international Florida artificial intelligence research society conference. 1998:235-239.

[25] Sebban M. On feature selection: A new filter model. In: Proceedings of the twelfth international flairs conference. 1999:230-234.

[26] Chuang LY, Ke CH, Yang CH. A hybrid both filter and wrapper feature selection method for microarray classification. In Proceedings of the international multi conference of engineers, and computer scientists. Hong kong: Imecs. 2008.

[27] Blessie EC, Karthikeyan E. Sigmis. A feature selection algorithm using correlation based method. J algor comp technol. 2012;6:385-394.

[28] Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374:20150202.

[29] https://vitalflux.com/feature-extraction-pca-python-example/

[30] Gabriel KR. The biplot graphical display of matrices with application to principal component analysis. Biometrika. 1971;58:453-467.

[31] http://archive.ics.uci.edu/ml

[32] https://www.codecademy.com/article/normalization

[33] https://www.investopedia.com/terms/m/multicollinearity.asp

[34] https://en.wikipedia.org/wiki/Principal_component_regression#:~:text=In%20statistics%2C%20principal%20component%20regression,a%20standard%20linear%20regression%20model

[35] Kennedy J, Eberhart R. Particle Swarm Optimization. In: Proceedings of the Ieee Int Conf Neural Network. 1995;4:1942-1948.

[36] Yeoman TB, Xue B, Zhang M. Particle Swarm Optimization for Feature Selection: A Size Controlled Approach. Proceedings of the 13th Australasian Data Mining Conferences on Research and Practice in Information Technology. 2015:151-159.

[37] Azevedo GL, Cavalcanti GDC, Carvalho Filho EC. An approach to feature selection for keystroke dynamics systems based on PSO and feature weighting. In: Proceedings of the 2007 congress on evolutionary computation. IEEE Publications. 2007:3577-3584.

[38] Pathak A, et al. Classification Rule and Exception Mining Using Nature Inspired Algorithms. Int J Comput Sci Inf Technol. 2015;6:3023-3030.

[39] Xue B, Zhang, Browne WN. Multiobjective Particle Swarm Optimization for Feature Selection. Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation. 2012:81-88.

[40] https://doc.lagout.org/science/Artificial%20Intelligence/Swarm%20Intelligence/Swarm%20intelligence%20-%20James%20Kennedy.pdf

[41] Agustian F, Lubis MD. Particle swarm optimization feature selection for breast cancer prediction. In 2020 8th International Conference on Cyber and IT Service Management. CITSM. IEEE. 2020:1-6.

[42] Tran B, Xue B, Zhang M. Overview of particle swarm optimization for feature selection in classification. Proceeding of 10th International Conference on Simulated Evolution and Learning. 2014:605-617.

[43] https://software.broadinstitute.org/morpheus/

[44] Han Z, Wei B, Zheng Y, Yin Y, Li K, et al. Breast Cancer Multi-Classification From Histopathological Images With Structured Deep Learning Model. Sci Rep. 2017;7:4172.

[45] Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial Intelligence in Digital Pathology – New Tools for Diagnosis and Precision Oncology. Nat Rev Clin Oncol. 2019;16:703-715.

[46] Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, et al. A Method for Normalizing Histology Slides for Quantitative Analysis. IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Boston, MA, USA 2009:1107-1110.

[47] Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, et al. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. JAMA. 2015;313:1122-1132.

[48] Madabhushi A, Lee G. Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities. Med Image Anal. 2016;33:170-1754.

[49] Sirinukunwattana K, Snead D, Epstein D, Aftab Z, Mujeeb I, et al. Novel Digital Signatures of Tissue Phenotypes for Predicting Distant Metastasis in Colorectal Cancer. Sci Rep. 2018;8:13692.

[50] Schmidt U, Weigert M, Broaddus C, Myers G. Cell detection with star-convex polygons. International Conference on Medical Image Computing and Computer-Assisted Intervention. Medical Image Computing and Computer Assisted Intervention (MICCAI). 2018.

[51] Ma J, Shao W, Ye H, Wang L, Wang H, et al. Arbitrary Oriented Scene Text Detection via Rotation Proposals. IEEE Trans Multimedia. 2018;20:3111-3122.

[52] Weigert M, Schmidt U, Haase R, Sugawara K, Myers G. Star-Convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. The IEEE Winter Conference on Applications of Computer Vision 2020:3666-3673.

[53] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). 2015:234-241.

[54] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Realtime Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:779-788.

[55] Ma J, Shao W, Ye H, Wang L, Wang H, et al. Arbitrary Oriented Scene Text Detection via Rotation Proposals. IEEE Trans Multimedia. 2018;20:3111-3122.

[56] https://github.com/stardist/stardist

Appendix A

Algorithm of PSO

Given:

An n-dimensional search space: $X = \{x1, x2, \ldots, xn\}$.

An objective function, $F : X\ R$ to be maximized.

Initialization:

Generate n particles, where $xi(0)\ X$ is position of particle i

For each particle, j:

Set **particle fitness** $q^i\ (0) = F\ (x^i\ (0))$

/* All particles are assigned fitness values and are evaluated by fitness functions to be optimized.

$b^i = f^i\ (0) \leftrightarrow$ **particle best quality** for i

$p^i = x^i\ (0) \leftrightarrow$ **particle best position** for i

Set **particle velocity** $v^i\ (0) = 0$

/* Particles have velocities that direct the flying of particles.

Set $b = \max\ (b^i) \leftrightarrow$ **global best quality**

Set $i^* = \text{argmax}\ (b^i) \leftrightarrow$ **global best particle**

Set $p = x^i * (0) \leftrightarrow$ **global best position**

For time t = 1 to t_{max}:

For each particle, i:

Update particle velocity:

$$vkj(t) = \omega vkj(t-1) + C\,rc(t)[pkj - xkj(t)] + S\,rs(t)[pk - xkj(t)]$$

Update particle position:

$$xkj(t+1) = xkj(t) + vkj(t).$$

Evaluate quality $fi(t) = F(xi(t))$.

if $fi(t) > bi\ bi = f(t)$.

/* Update particle's best position if needed

$$pi = xi(t).$$

if $\max(bi) > bb = \max(bi)$.

/* Update global best position if needed $i* = \text{argmax}(bi)$. and $p = xi * (t)$

The PSO algorithm employs particles, each with its own parameter values that vary based on recent experience. This approach is simple, efficient, effective, and easy to implement, making it applicable to a broad range of problems. Its flexibility enables it to solve difficult problems with ease.

Appendix B

U-Net

A typical architecture of a convolutional neural network (CNN) comprises of a sequence of layers wherein the input layer is the image that is to be processed. The convolutional layer has several sub-layers each representing a filter mask that is convolved with the entire image in parallel. Each filter mask is typically smaller than the input image. Pixels are convolved using a filter or a kernel to create feature maps. The result of this step is a dot product between a patch and the kernel. In the next step, subsampling is performed by using either Max, Min, or Average pooling types. The result of this step is to reduce the data dimensionality and to reduce overfitting. Using a sequence of convolution and pooling layers, the output is then fed to a fully connected layer for classification.

Convolutional networks typically make use of classification tasks where the image output is a single class label. However, in the context of biomedical image processing, where thousands of training images are beyond reach, it becomes important that the desired output should contain localization particularly when an image consists of many visual tasks, and in such cases, it is required that a class label is assigned to each pixel.

U-Net architecture proposed in [53], comprises of a contracting and symmetric expanding paths wherein contracting path is used to capture the context and the latter is used to provide a precise localization. The architecture of U-Net is illustrated in FIGURE 17. The contracting path makes use of the typical architecture of a convolutional network. It comprises of the repeated application of two 3x3 convolutions that are unpadded, with each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for down sampling. The number of feature channels are

doubled at each down sampling step. While in the expansive path, each step follows the up sampling of the feature map followed by a 2x2 up convolution, that halves the number of feature channels, and a concatenation with the corresponding cropped feature map of the contracting path, and two 3x3 convolutions, with each followed by a ReLU. This cropping feature is necessary because of the loss of border pixels in every convolution. Finally, at the last layer a 1x1 convolution is used that maps each 64- component feature vector to the desired number of classes. On the contrary, the total network has 23 convolutional layers.

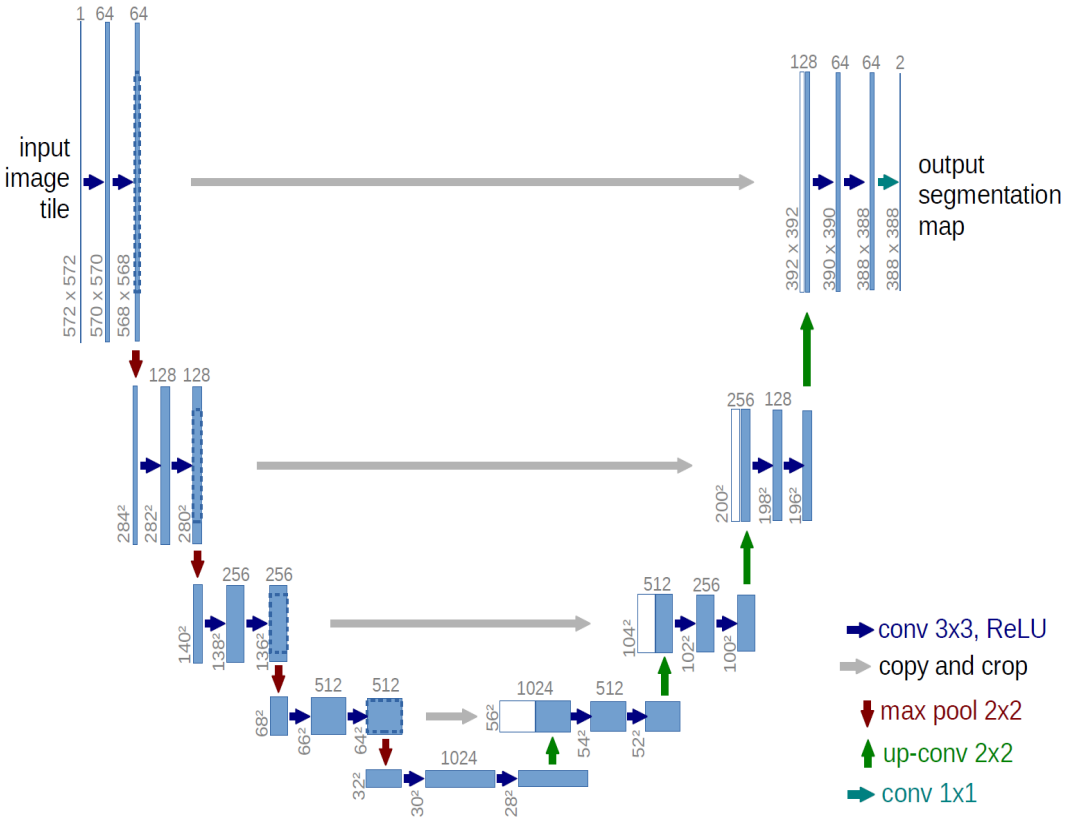


Figure 17: Architecture of U-Net (32x32 pixels in the lowest resolution) [53].

As seen from the figure above, each box in blue corresponds to a multi-channel feature map. The number of channels is presented at the top of each box. The size of x-y is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows representing different colors exhibit different operations.