10.48047/jocaaa.2023.31.04.59

Enhancing Breast Cancer Diagnosis with Multivariate Bayesian PCA Model using Machine learning

P. Srivyshnavi[1], M Darshan Teja [2*] ,P Shankaraiah [3], Sreenivasulu T [4]

[1]Assistant Professor , Dept. of CS & E, S.P.M.V.V. Engineering College, Tirupati, India.
[2, 3, 4] Department of Mathematics, School of Advance Science, VIT Vellore, India.

Corresponding author: darshanteja49@gmail.com *

Abstract:

This research presents a multivariate Bayesian PCA model as a potential method for more accurately diagnosing breast cancer. The model uses a comprehensive dataset that includes attributes such as "radius", "texture", "perimeter", and "area" across dimensions (including "mean," "SE," and "worst") in order to forecast whether or not the outcomes will be "malignant" or "benign. Principal Component Analysis (PCA), which is followed by a multivariate Bayesian technique using a Gaussian Naive Bayes classifier for diagnosis prediction, simplifies difficult problems while preserving essential information. The findings demonstrate that the model is effective, F1-score is 0.94, precision score is 0.93, recall score is 0.94, and model's accuracy is 0.92, respectively. AUC of 0.91 denotes the area under the curve is produced by the receiver operating characteristic (ROC) analysis, which demonstrates excellent classification capabilities. The insights provided by computed covariance matrices offer a greater understanding of the relationships between attribute values. In conclusion, the combination of principal component analysis with Bayesian improves breast cancer diagnosis in terms of both accuracy and efficiency; however, this improvement is contingent on clinical validation before it can be applied in the real world.

Keywords: Breast Cancer Diagnosis, MBPCA, GNB, Accuracy, Attribute Insights.

1. Introduction

Breast cancer is a severe problem worldwide and impacts people of all ages, races, and socioeconomic backgrounds. It has a great impact on public health systems since it is the cancer with the highest incidence rate in females. This cancer starts in the breast and can present in several distinct ways, most often through the milk ducts or lobules. Early detection and effective treatment are extremely important to successfully fight breast cancer and improve patient outcomes. Breast cancer is a complex disease with many possible causes including genetics, the environment and the habits and lifestyle choices of an individual. Breast cancer is a complex disease that has evolved into a myriad of subtypes, each with its unique set of symptoms and treatment response. The best treatment effect can be reached by timely and precise diagnosis, which allows tailoring the interventions to the certain subtype and stage of the disease.

Breast cancer diagnosis is based on a wide array of traits that are important markers of the underlying pathology. These include the properties "radius", "texture", "perimeter", "area",

10.48047/jocaaa.2023.31.04.59

"smoothness", "compactness", "concavity", "convexity", "symmetry" and "fractal dimensions". Note that these qualities are quantified using mean, standard error (SE) and worst values. The two together can provide insights regarding tumor size, textural patterns, structural features, and more. In this case, the power of breast cancer diagnosis can be improved by combining Gaussian Naive Bayes (GNB) with PCA or Principal Component Analysis. PCA reduces the dimensionality of the complex web of features while preserving the integrity of the most important information. To conclude PCA is extended to the Bayesian framework, which is a probabilistic classification procedure in GNB. This approach gives a numerical measure of uncertainty that can assist in diagnostic inferences. This combination method does not limit to binary categorization, as traditional regression models do, and so offers nuanced insights into the possibility of malignancy. This allows more educated and individualized therapy treatments.

In this article, we'll delve into the complex world of breast cancer and shed light on the characteristics that are crucial to making a diagnosis. We also reveal the promise of multivariate Bayesian PCA regression with GNB for better comprehending and treating this complicated illness. This combination method paves the way for in-depth research into its diagnostic capabilities, which has great promise for improving breast cancer diagnosis accuracy and efficiency have reached new heights.

2. Literature Review

Enhancing the methods used for feature extraction and dimensionality reduction was the main goal of Nounou et al.'s (2002), [1]seminal research. This was achieved by introducing Bayesian Principal Component Analysis (BPCA) as a novel approach. The researchers undertook a transformational endeavor by integrating key equations to produce the Maximum Likelihood Principal Component Analysis (MLPCA) approach. The unique methodology demonstrated the capacity for improved data analysis and pattern detection. After that, the authors used these new ways to look at old data, which helped them show the benefits and possibilities of BPCA, MLPCA, and traditional PCA models. The research conducted by the authors emphasized the adaptability of Bayesian-inspired methodologies and their capacity for wider utilization beyond real-time data. This study has established a strong basis for future progress in the fields of statistical modeling and machine learning.

Tsehay and Sushma [2] initiated an endeavor to leverage Machine-Learning techniques for the prediction of breast cancer, utilizing two robust models, namely SVM and Decision Trees, as their analytical tools. The study yielded significant findings, as the SVM had a notable Accuracy rate is 91.92%. This outcome highlights the SVM's ability to effectively differentiate between malignant and benign cases. In the context of this essential diagnostic work, the Decision Tree model exhibited its efficacy as a beneficial tool, but with a slightly lower accuracy rate of 87.12%. Their findings highlight the promise in the use of machine learning algorithms enhancing the breast cancer prediction & providing physicians and researchers with vital tools for early identification and well-informed medical interventions.

Within the domain of breast cancer diagnosis, a multitude of research endeavors have been undertaken to investigate the effectiveness of different [3] machine learning algorithms. These

10.48047/jocaaa.2023.31.04.59

endeavors have yielded significant contributions towards enhancing the diagnostic procedure. A study was conducted to investigate the effectiveness of KNN or K-Nearest Neighbors [4], Random Forest (RF), & Naive Bayes (NB) algorithms. Among these algorithms, KNN demonstrated superior performance by consistently achieving accurate classifications, highlighting its efficacy in this regard. In a separate study undertaking, the [2], [5]decision tree model emerged as a prominent method, showcasing its efficacy in the detection of breast cancer.

In addition, it's worth noting that the utilization of an ANN or artificial neural network [6] model demonstrated exceptional efficacy in the analysis of Wisconsin's breast cancer dataset, thereby emphasizing the wide-ranging capabilities of neural network-based methodologies[7]. In the meantime, an independent study explored the field of unsupervised learning, utilizing K-means clustering to attain a noteworthy accuracy rate of 73.70% [8]. The results presented in this study highlight the wide range of machine learning model that can be used to forecast breast cancer. These techniques include [9] "Naive Bayes (NB)", "Deep learning (DL)", "support vector machines (SVM)" [2], [10] and others[11]. Each of these approaches has its own unique strengths and advantages, which can significantly contribute to the ongoing efforts to combat this prevalent disease

3. PCA and Bayesians

3.1 Principal Component Analysis

PCA serves to represent the original matrix in relation to transformed variables and novel axes of rotation, of then referred to as loadings or projection directions. This transformation is achieved through a matrix multiplication operation. [1], [12]Given an initial matrix of process variables with dimensions **m x n**, denoted as $\boldsymbol{X} = \widetilde{\boldsymbol{X}} + \boldsymbol{\varepsilon_x}$, where $\widetilde{\boldsymbol{X}}$ represents the matrix capturing the intrinsic noise free information, and $\boldsymbol{\varepsilon_x}$ is the additive noise matrix. Here, **n** represents the number of variables, and **m** represent the number of observations, PCA's breaks it down into its component parts as **X**

$$\boldsymbol{X} = \boldsymbol{Y\theta^T} \tag{1}$$

In this case, the principal components (or PCs) are represented by a m x n matrix (Y), and the loadings (or projection directions) are represented by an orthogonal n x n matrix θ. The data covariance matrix $\boldsymbol{X^T X}$, is transformed through diagonalization.

Where $$\boldsymbol{X^T X} = \boldsymbol{\theta M \theta^T} \tag{2}$$

where **M** is an orthogonal matrix holding the covariance matrix of data is eigenvalues. When Eq'n (1) is substituted into Eq'n (2), we get

$$\boldsymbol{X^T X} = (\boldsymbol{Y\theta^T})^{\boldsymbol{T}}(\boldsymbol{Y\theta^T}) = \boldsymbol{\theta Y^T Y \theta^T} = \boldsymbol{\theta M \theta^T} \tag{3}$$

It shows that the PCs are independent variables whose variances are the same as the eigenvalues of the covariance matrix that describes the data.

10.48047/jocaaa.2023.31.04.59

The first component of the PCA can be estimated by solving the following optimization problem:

$$\{\widehat{\boldsymbol{\theta_1}}, \widehat{\boldsymbol{Y_1}}\}_{PCA} = \boldsymbol{argmax}_{\widehat{\theta_1}, \widehat{Y_1}} \{\boldsymbol{var}(\boldsymbol{X}\widehat{\boldsymbol{\theta_1}})\} \quad (4)$$

In order to minimize the total estimation error, we can recast the PCA estimation issue depicted in equation 4 as the following optimization problem:

$$\{\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{Y_\iota}}\}_{PCA} = \boldsymbol{argmin}_{\widehat{\theta}, \widehat{Y_\iota}} \sum_{i=1}^{n} (\boldsymbol{X_i} - \widehat{\boldsymbol{X_\iota}})^T (\boldsymbol{X_i} - \widehat{\boldsymbol{X_\iota}}) \quad (5a)$$

$$s.t. \widehat{\boldsymbol{X_\iota}} = \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{Y_\iota}} \; \boldsymbol{and} \; \widehat{\boldsymbol{\theta}}^T\widehat{\boldsymbol{\theta}} = \boldsymbol{1} \quad (5b)$$

$\widehat{\boldsymbol{Y_\iota}}$ a p x 1 vector for [12] estimated main observational component is $\boldsymbol{X_i}$, and $\boldsymbol{X_i}$ and $\widehat{\boldsymbol{X_\iota}}$ are n x 1 vectors for the ith set of observed and estimated observations, respectively. These vectors, **X** and Y, represent the transposed rows of their respective matrices for notational purposes. Since equation (5a) uses an identity normalization matrix, it can be deduced that PCA assumes that all variables have the same amount of noise.

3.2 Introduction to Bayesian

The peculiar aspect of Bayesian estimation is that it regards all values, observable and unobservable, as random and assigns them a probability density function that reflects their joint behaviour. Unlike many non-Bayesian approaches that regard the quantity of items of interest as constant unknowns, Bayesian methods [13]–[15] regard them as random variables which can be searched by minimizing a particular objectives function of estimating errors. This is a presupposition for Bayesian techniques and it is feasible to incorporate outside prior knowledge about the quantities of interest. In Bayesian estimation, the first step is to calculate $\boldsymbol{P(E_i|A_i)}$, given the measurements, the conditional density with respect to the variable to be approximated also known as Posterior. This is done so that the quantity E_i can be calculated from a sequence of quantity measurements $\boldsymbol{A_i}$.

The posterior variable is a density function that characterizes how the quantity $\boldsymbol{E_i}$ is expected to behave in light of the observed data. The posterior can be expressed in standard form using Bayes' rule:

$$\boldsymbol{P(E_i|A_i)} = \frac{\boldsymbol{P(A_i|E_i)P(E_i)}}{\boldsymbol{P(A_i)}} \quad (6)$$

Equation (6)'s first numerator term is the probability function, This refers to a series of observations under the assumption that $\boldsymbol{E_i}$ is really what it is supposed to be. The likelihood principle states that the probability functions takes into consideration each and every one of the features supplied by the observations $\boldsymbol{A}$ regarding the quantity $\boldsymbol{E_i}$. [1]Prior, defined as the density function of the quantity $\boldsymbol{E_i}$, is the second term in the numerator. A previous measurement is a measure of how much we thought we knew about $\boldsymbol{E_i}$ before we actually took the measurement.

10.48047/jocaaa.2023.31.04.59

The estimate problem is improved by incorporating the prior, which is outside information on quantity $\boldsymbol{E_i}$. Finally, the density function of the observation serves as the denominator term because it is believed to be constant after the data observation. This allows us to express the posterior density as

$$\boldsymbol{P(E_i|A_i) \propto P(A_i|E_i)P(E_i)} \quad (7)$$

the un-normalized posterior, as it is called in full. Hence, the posterior consists of internal and external elements. Once the posterior has been constructed, a subset of it is chosen as the final Bayesian estimate. Unlike frequentist or non-bayesian approaches which employ only the data to derive conclusions, bayesian methods incorporate both the data and a part of previous understanding represented by the prior, improving the quality of estimation.

3.3 Loss function

Loss function, a untility function $L(E_i, \widehat{E_i})$, determines that fact posterior representative piece will serve as the basis for the bayesian estimate. Bayesian estimates of E (denoted by E_i) and the actual value of E (denoted by $\widehat{E_i}$) are shown here i.e.,

$$\mathbf{L(E_i, \widehat{E_i})}=\begin{cases}\mathbf{0} & \boldsymbol{when}\ \{\widehat{\boldsymbol{E_i}}\}_{\boldsymbol{Bayesian}} = \boldsymbol{E_i} \\ \mathbf{1} & \boldsymbol{otherwise}\end{cases} \quad \mathbf{(8)}$$

The maximum a posteriori (MAP) estimate is the Bayesian estimate most commonly associated with the usage of a zero-one loss function. Thus

$$\{\widehat{\boldsymbol{E_i}}\}_{\boldsymbol{MAP}} = \boldsymbol{argmax}_{\widehat{E_i}} \boldsymbol{P(A_i|E_i)P(E_i)} \quad (9)$$

The BPCA Algorithm was created by employing a loss function of zero and one. The use of this loss function simplifies the BPCA comparison process by reducing the Bayesian PCA modeling to a minimization problem.

3.4 Bayesian principal component Analysis (BPCA)

To define the PCA model using a data matrix, it is essential to estimate the projection directions, the principle components and the actual modelling rank. Hence, the posterior in a Bayesian construction is to be interpreted as the conditional density of such values given the data observed. This can be stated using the bayes rule as

$$\boldsymbol{P(\widetilde{Y}, \widetilde{\theta}, \widetilde{p}|X)} = \frac{\boldsymbol{P(X|\widetilde{Y}, \widetilde{\theta}, \widetilde{p})P(\widetilde{Y}, \theta, \widetilde{p})}}{\boldsymbol{P(X)}} \quad \mathbf{(10)}$$

The conditional density of the observed variables stated the noise-tree PCA model and the data is the likelihood function, and the prior is the second term in the numerator. One possible expression for the unnormalized posterior is

10.48047/jocaaa.2023.31.04.59

$$\boldsymbol{P}(\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{p}}|\boldsymbol{X}) \propto \quad \boldsymbol{P}(\boldsymbol{X}|\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{p}})\boldsymbol{P}(\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{p}}) \tag{11}$$

The prior is a extremely involved function that represents the combined density of the noise-free principal components, projection instruction, and the genuine PCA model. However, model rank affects both the principal component density function and the projection direction. Hence, the preceding can be expressed as [16]

$$\boldsymbol{P}(\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{p}}) = \boldsymbol{P}(\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{p}})\boldsymbol{P}(\widetilde{\boldsymbol{p}}) \tag{12}$$

Additionally, using the multiplication rule of probabilities, the PCA main component joint density function and projections instruction may be represented as

$$\boldsymbol{P}(\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{p}}) = \boldsymbol{P}(\widetilde{\boldsymbol{Y}}|\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{p}})\boldsymbol{P}(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{p}}) \tag{13}$$

3.5 Loss function

In this particular piece of research, a 0-1 loss function of the frame is utilized.

$$\boldsymbol{L}(\widehat{\boldsymbol{Y}},\widehat{\boldsymbol{\theta}};\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}}) = \begin{cases} \boldsymbol{0} & \boldsymbol{when}\ \{\widehat{\boldsymbol{Y}},\widehat{\boldsymbol{\theta}}\}_{\boldsymbol{Bayesian}} = \widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}} \\ \boldsymbol{1} & \boldsymbol{other\ wise} \end{cases} \tag{14}$$

Consequently, the solution to the BPCA is able to achieved by addressing the optimization's problem that follows:

$$\{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{Y}}\}_{\boldsymbol{Bayesian}} = \boldsymbol{argmax}_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{Y}}}\boldsymbol{P}(\boldsymbol{X}|\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{p}})\boldsymbol{P}(\widetilde{\boldsymbol{Y}}|\widetilde{\boldsymbol{\theta}})\boldsymbol{P}(\widetilde{\boldsymbol{\theta}}) \tag{15}$$

formulation, as an example, generates a solution in closed form for the predicted data, [17] which is both computationally highly effective and permits immediate comparison to other approaches like principal component analysis.

3.6 Gaussian Navie Bayes

The Naive Bayes classification approach assumes that each feature that predicts the target value is independent of the others according to Bayes' Theorem. It works out the odds for each group and picks the one with the highest odds. It is good for many things but one area where it particularly excels is handling difficulties related to Natural Language Processing (NLP). Naive Bayes Classifier assumes that the features used for making the target prediction have no effect on each other. Unlike real world data where features are dependent on each other for determining the target, Naive Bayes classifier does not take this into consideration. The assumption of independence is always erroneous for real data, but is useful in many cases. which we call "Naive".

From equation (6) Bayes' theorem states, given a feature vector $\boldsymbol{A_i} = (\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots\ldots\ldots \boldsymbol{a_n})$ and a class variable E_i, Our goal is to get $\boldsymbol{P}(\boldsymbol{E_i}|\boldsymbol{A_i})$ given the likelihood $\boldsymbol{P}(\boldsymbol{A_i}|\boldsymbol{E_i})$ and prior probabilities

10.48047/jocaaa.2023.31.04.59

$\boldsymbol{P(E_i)}$ and $\boldsymbol{P(A_i)}$.
It is possible to break down the probability $P(A_i|E_i)$ using the chain rule as follows:

$$\boldsymbol{P(A_i|E_i) = P(a_1, a_2, \dots\dots\dots\dots\dots, a_n|E_i)} \quad \textbf{(16)}$$

$$\boldsymbol{P(A_i|E_i) = P(a_1|a_2, \dots\dots, a_n, E_i)P(a_2|a_3, \dots\dots, a_n, E_i) \dots\dots\dots P(a_n|E_i)} \quad \textbf{(17)}$$

However, the conditional probabilities are not interdependent due to the naive conditional independence requirement.

$$\boldsymbol{P(A_i|E_i) = P(a_1|E_i)P(a_2|E_i) \dots\dots\dots\dots\dots P(a_n|E_i)} \quad \textbf{(18)}$$

So, using this definition of conditional independence:

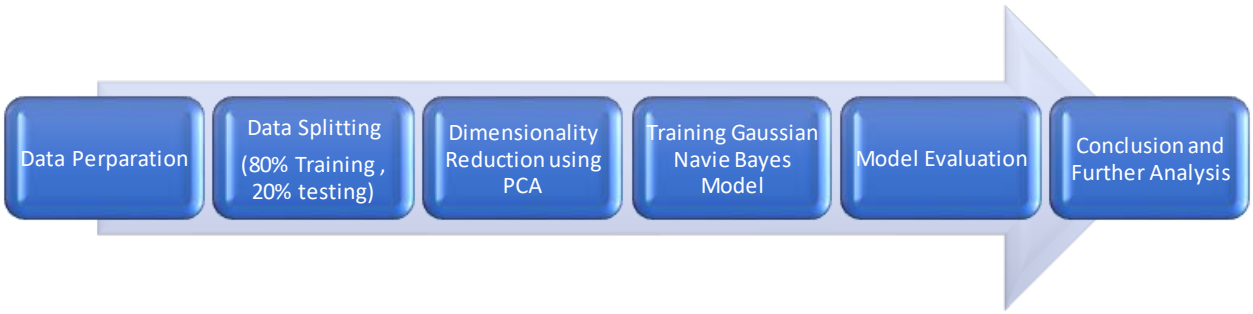
$$\boldsymbol{P(E_i|A_i) = \frac{P(a_1|E_i)P(a_2|E_i)\dots\dots\dots P(a_n|E_i)P(E_i)}{P(a_1)P(a_2)\dots\dots\dots P(a_n)}} \quad \textbf{(19)}$$

With a fixed denominator, the posterior probability can be written as:

$$\boldsymbol{P(E_i|a_1, a_2, \dots\dots\dots, a_n) \propto P(E_i)\prod_{i=1}^{n} P(A_i|E_i)} \quad \textbf{(20)}$$

The combination of Machine Learning with PCA and Gaussian Naive Bayes is a powerful one, and demonstrates the efficiency and versatility of Python and R. PCA reduces data dimensions while retaining significant information, improving the efficiency and accuracy of analysis. Manual analysis of high-dimensional data is not easy. PCA makes feature extraction and reduction easier by automatically identifying relevant data patterns and structures. PCA reduces our breast cancer dataset to the most important features and discards the others. This reduces the data and increases the performance and the readability of Gaussian Naive Bayes classifier. The key is the accuracy, F1-score and other evaluation metrics of this integrated technique. We use machine learning to improve the accuracy, efficiency, and availability of breast cancer diagnosis using PCA and Gaussian Naive Bayes. This synergistic effort demonstrates how automation can address difficult problems and how data driven healthcare may function.

4 Methodology



10.48047/jocaaa.2023.31.04.59

Fig-1: Methodology for MBPCA

According to the aforementioned Fig(1), the algorithm for predicting breast cancer entails the process of data preparation, which includes loading the information and encoding the diagnosis labels. Subsequently, [6] the data is split into two groups, one for training and one for testing, with 80% among the input going to the former and 20% to the latter or into the set used for testing. Principal Component Analysis (PCA) is commonly employed to decrease dimensionality and identify essential information, typically preserving a set of eight principal components. Subsequently, the training data that has been converted using principal component analysis (PCA) is utilized to train a Gaussian Naive Bayes classifier. This classifier is able to discern patterns inside the reduced-dimensional space.

The evaluation of the model's performance is conducted comprehensively by employing various measures such as Accuracy, recall, precision, F1-score, & ROC_AUC. These Measure are then visually represented through the utilization of a confusion matrix and ROC curve. The investigation yields findings regarding the predictive ability of the model for breast cancer. Subsequent investigations may delve into the significance of features, optimization of hyperparameters, implementation of cross-validation, and other augmentations in an effort to propel the reliability and exactness of the predictive model.

5 Analysis

We used multivariate Bayesian principal component analysis (MBPCA) to investigate 570 breast cancer (Kaggle data set) cases with variables categorized as "mean," "SE," and "worst." This innovative method combines PCA and Gaussian Naive Bayes to reduce and classify dimensions in a shared framework. Our research began with MBPCA, which reduces data dimensionality without losing useful information. Making the analysis simpler enhances breast cancer diagnosis. We next examined the findings to evaluate our model and found a correlation matrix that illuminated our variables' relationships. This matrix showed variable relationships, revealing the data structure. The confusion matrix, Accuracy's,[8] recall, precision, F1-Score & ROC_AUC of the MBPCA model were also evaluated. The model's accuracy, F1-scores, and ROC AUC values were highlighted in this detailed analysis. Our research shows the MBPCA model's breast cancer detection potential. Our PCA-Gaussian Naive Bayes approach improves dimensionality reduction and predictive analytics. Our findings suggest improved patient care and customized treatment regimens, which could improve breast cancer diagnosis.

5.1 Correlation Matrix

As part of our exploratory study of breast cancer data, we took the time to carefully look at the connections between a number of key variables. We looked at a wide variety of characteristics, including the[18] "Radius-mean", "Texture-mean", "Perimeter-mean", "Area-mean", "Smoothness-mean", "Compactness-mean", "Concavity-mean", "Concave points-mean", "Symmetry-mean", "Fractal Dimension-mean", "Radius-se", "Texture-se", "Perimeter-se", "Area-se", etc. by conducting this in-depth analysis, we were able to develop a correlation matrix that sheds light on the statistical connections between the aforementioned factors. The correlation

10.48047/jocaaa.2023.31.04.59

matrix provided a graphical and numerical representation of the direction and intensity of correlations, thereby revealing interdependencies and patterns in the data.

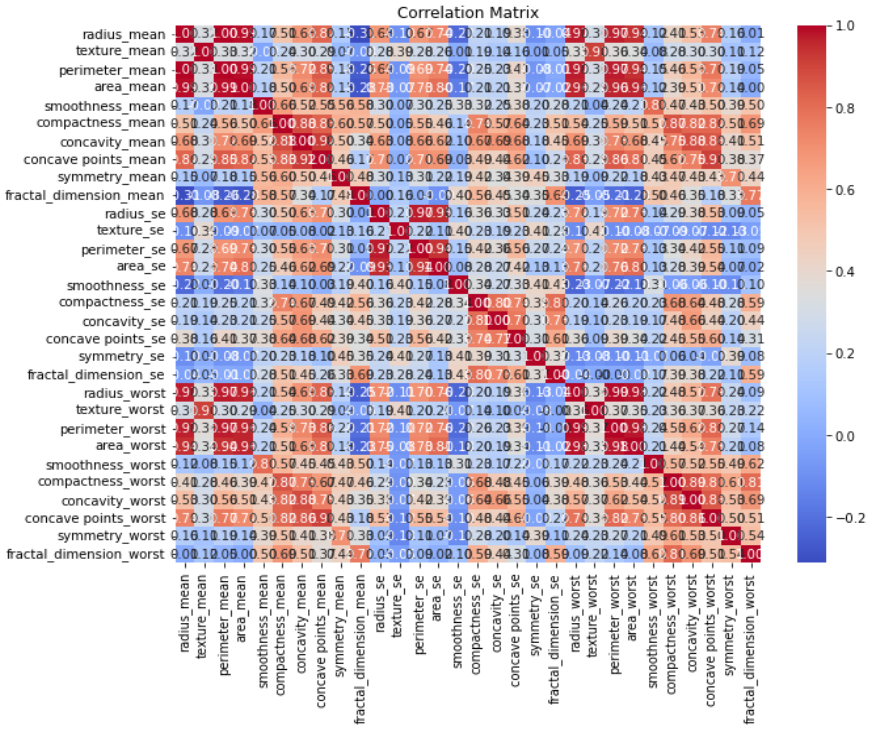


Fig-2: Correlation Matrix for Breast cancer

This study enabled us to explore potential interdependencies and correlations between these critical features, clarifying which characteristics may be more relevant in breast cancer detection. Recognition of these relationships is critical to uncover the underlying disease and to guide research and therapy decisions moving forward. In summary, the correlation matrix we constructed can be a useful tool to clarify the complicated interrelationships among the numerous components in our breast cancer dataset. These discoveries may help both doctors and patients build more accurate diagnostic and prognostic models in the battle against breast cancer.

5.2 Confusion Matrix

10.48047/jocaaa.2023.31.04.59

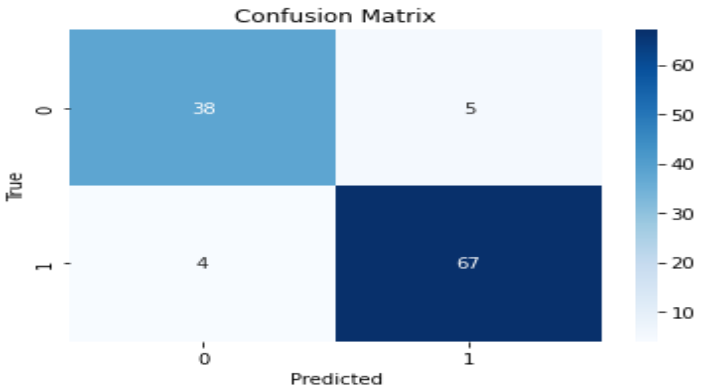


Fig-3

We thoroughly analyzed breast cancer diagnosis models using the Confusion Matrix, a standard classification accuracy measure. We used 80% of our data for model training and 20% for testing, per machine learning standards. In this paradigm, we categorized our diagnosis labels as malignant ('M' = 0) or benign ('B' = 1). The confusion matrix demonstrated the model's class distinction. The matrix element (0,0) indicates the 38 times the model recognized malignant patients ('M'). The element (0,1) denoted the 5 times the model misidentified benign cases ('B') as malignant. Four instances where the model misidentified malignant cases as benign are (1,0). Finally, (1,1) indicated occurrences where the model properly recognized benign cases as benign. This number was high, at 67. These findings show that the model accurately detects malignancies and minimizes benign false positives. However, the (0,1) and (1,0) components demonstrate the importance of reducing false positives. This detailed analysis of the Confusion Matrix improves breast cancer diagnosis and patient care by optimizing the model.

5.3 Accuracy, Precision, Recall, F1-Score

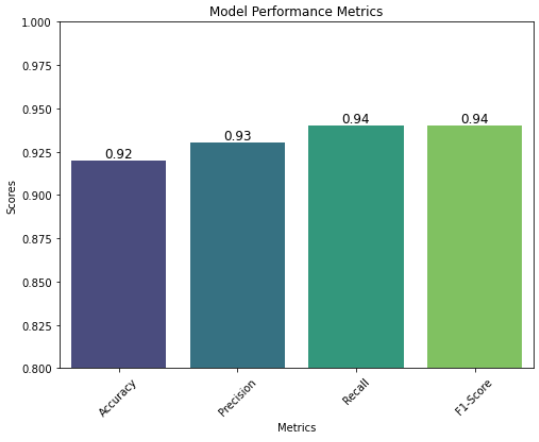


Fig- 4: Model perform metrics

10.48047/jocaaa.2023.31.04.59

As part of our effort to better diagnose breast cancer, we turned to multivariate Bayesian principal component analysis (MBPCA). With this novel method, we were able to evaluate the performance of our model with striking success. Our MBPCA model, which deftly combines PCA and Gaussian Naive Bayes, produced impressive results. Vital diagnostic measures are the focus of our investigation. With an accuracy of 0.92, our model demonstrated impressive precision in its classifications. Incredibly, the F1 Score, a crucial metric that strikes a balance between precision and recall, was 0.94.

This suggested that the model's sensitivity to detecting malignancy was well-balanced with its accuracy in making positive predictions, providing a useful resource for doctors. Our model's ability to reduce false positives while correctly identifying malignant patients was demonstrated by its high recall and precision scores. Our MBPCA model successfully distinguished between benign and malignant patients with 0.93 precision and a 0.94 recall.

These findings not only show promise but also have substantial potential to improve breast cancer diagnosis. Our MBPCA model is suitable for real-world clinical applications because of its accuracy and balance between precision and recall. In the ongoing fight against breast cancer, our findings pave the path for better patient care, faster interventions, and better-informed medical decisions.

5.4 Receiver Operating Characteristic

When evaluating the efficacy of a classification model, in particular when examining the process of diagnosing breast cancer, the Receiver Operating Characteristic Area Under the Curve or (ROC_AUC) is an essential statistic. As you can see in the accompanying image, our research yielded an impressive ROC AUC score of 0.91. This remarkable result demonstrates how well our approach can distinguish between cancerous and benign conditions.

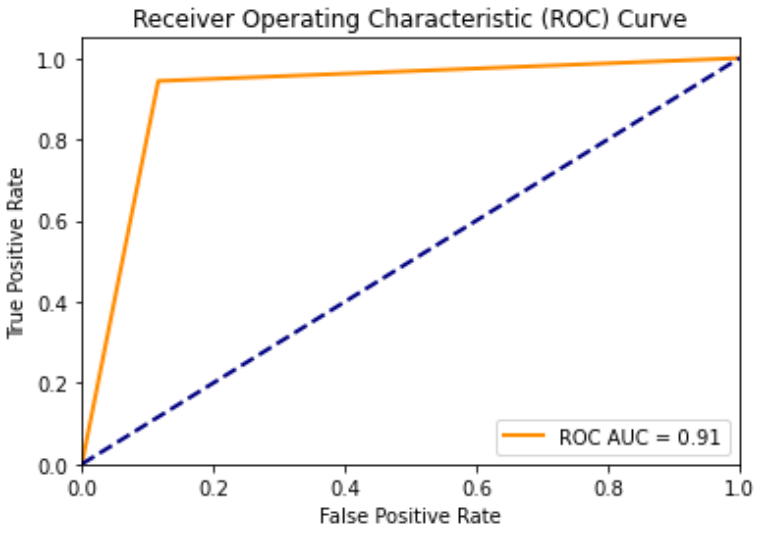


Fig-5: ROC_AUC

10.48047/jocaaa.2023.31.04.59

The ROC AUC curve graphically illustrates the model's robustness in handling multiple diagnostic settings by depicting its ability to correctly categorize examples across varying thresholds. When the ROC AUC value is close to 1, it indicates that the model is very accurate by striking a good balance between the True Positive (TP) and False Positive (FP) Rates. The outstanding ROC-AUC of 0.91 for our breast cancer diagnostic model further demonstrates its dependability and therapeutic promise. This demonstrates the model's ability to provide reliable diagnoses, which can lead to better care delivered faster. Our results in this area have the potential to significantly enhance the field of breast cancer diagnostics and improve patient outcomes.

5.5 Scatter plots

In order to enhance our comprehension of the interconnections among different attributes within our breast cancer dataset, we utilized scatter plots as an effective means of visual representation. Our study specifically examined pairs of fundamental characteristics, that's like "Radius-mean", "Texture-mean", "Perimeter-mean", and "Area-mean", among other features. These features encompass essential data regarding the properties of breast tissue, offering useful insights for the purpose of diagnosis.

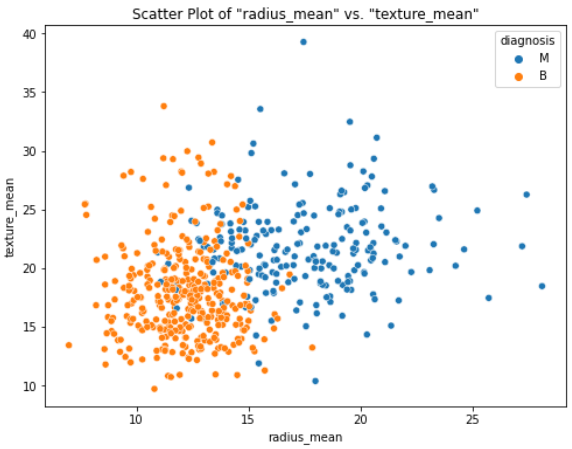


Fig-6 Radius vs Texture

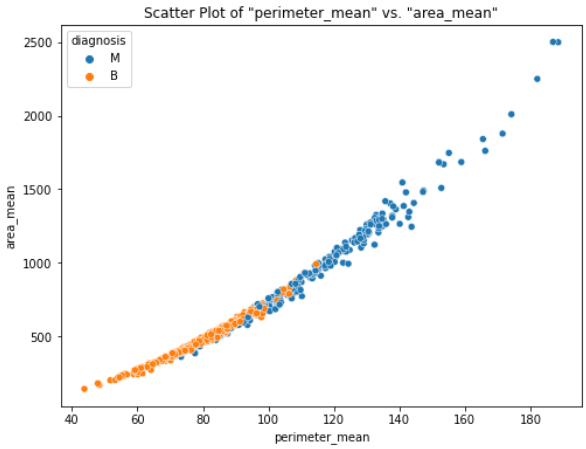


Fig-7 Perimeter vs Area

To provide visual examples, we give a selection of scatter plots displaying the variables 'radius mean' and 'texture mean,' as well as 'perimeter mean' and 'area mean.' The scatter plots effectively illustrate the interplay between different qualities, highlighting possible trends and associations. The scatter plot depicting the relationship between the 'radius mean' and 'texture mean' variables allows for the examination of the association between tumor size ('radius mean') and tissue texture ('texture mean'). In the plot that compares the variables "perimeter mean" and "area mean," we look for a link between the tumor's edge, which is shown by "perimeter mean," and its overall size, which is shown by "area mean."

The scatter plots presented offer a first glimpse into the underlying structure of the data, suggesting possible patterns and relationships that can be further explored and revealed by our analytical models. Although we have only shown a limited number of attribute pairings, our extensive

10.48047/jocaaa.2023.31.04.59

analysis encompasses numerous comparable visual representations, facilitating a full examination of the dataset. Visual insights contribute significantly to directing the selection of features and the creation of models, with the ultimate goal of achieving more accurate breast cancer detection and enhancing patient care. These insights are considered vital in this pursuit.

5.6 Cross Validation, Hyper Parameter

In our endeavor to enhance the precision of our breast cancer diagnostic model, we engaged in an exploration of hyperparameter optimization, employing the concepts of cross-validation as our guiding framework. With rigorous experimentation and thorough evaluation, we were able to attain a noteworthy cross-validation accuracy of 0.94, accompanied by a small confidence interval of +/- 0.03. The obtained outcome indicates a notable degree of stability and uniformity in the predictive capabilities of our model when applied to various subsets of the dataset.

The pursuit of ideal hyperparameters has resulted in the identification of the 'var smoothing' hyperparameter, which played a crucial role in improving the accuracy of our model. The setup that achieved the highest performance, when the 'var smoothing' parameter was set to 1e-09, resulted in an impressive accuracy score of 0.9385. The process of fine-tuning hyperparameters exemplifies our commitment to attaining the utmost level of diagnostic accuracy in breast cancer staging. The aforementioned results highlight the significance of hyperparameter adjustment in the construction of resilient machine learning models.

The utilization of rigorous cross-validation techniques, along with the determination of optimal hyperparameters, enhances the capability of our model to continuously produce precise and dependable outcomes. This accomplishment signifies a noteworthy advancement in our endeavor to boost the identification of breast cancer, so it augments the clinical effectiveness of the model and has the capacity to potentially improve patient care results.

6 Conclusion and Future Development

In conclusion, the study of breast cancer diagnostics using cutting-edge machine learning approaches has yielded optimistic results and noteworthy insights. Multivariate Bayesian Principal Component Analysis (MBPCA) is a groundbreaking method that combines Principal Component Analysis (PCA) with Gaussian Naive Bayes, and it has been used to increase the accuracy and efficiency of breast cancer categorization. We conducted a thorough analysis, and as a result, we have a full suite of performance metrics, including a great ROC AUC value is 0.91(91%), Precision value is 0.93 (93%), Recall value is 0.94 (94%), and an Accuracy rate is 0.92 (92%). These metrics highlight our model's reliability and medical viability in distinguishing between cancerous and noncancerous cases, an essential task for the prompt identification and implementation of therapeutic measures. We also included visual representations in our study results, such as scatter plots and correlation matrices, which provided additional insights into the interrelationships of features and the underlying structure of the data. Using hyperparameter tuning, cross-validation, and the identification of optimal hyperparameters, we were able to further enhance our model's precision and generalization capabilities.

10.48047/jocaaa.2023.31.04.59

We expect several promising avenues to be investigated as we plan for future developments. The model's applicability and accuracy could improve with the addition of new variables and data sets. It is possible that diagnostic accuracy could be improved through the use of deep learning and ensemble modeling. SHAP (Shapley Additive Explanations) values are one method that can be used to increase the model's interpretability and explainability. This helps doctors better comprehend the reasoning behind the model's conclusions. Implementation of the model in clinical settings is an important next step, as is continuous improvement via feedback loops. Breast cancer diagnosis may undergo a dramatic shift once such technology is implemented, leading to better accuracy, faster diagnosis, and more tailored care for each patient.

7. References

[1] Nounou MN, Bakshi BR, Goel PK, Shen X. Bayesian principal component analysis. Journal of Chemometrics: A Journal of the Chemometrics Society. 2002 Nov;16(11):576-95.

[2] Tsehay Admassu Assegie SS. A support vector machine and decision tree based breast cancer prediction. International Journal of Engineering and Advanced Technology (IJEAT), ISSN. 2020 Feb:2249-8958.

[3] Sindhwani N, Rana A, Chaudhary A. Breast cancer detection using machine learning algorithms. In2021 9th International conference on reliability, Infocom technologies and optimization (trends and future directions)(ICRITO) 2021 Sep 3 (pp. 1-5). IEEE..

[4] Ait Saadi T, Benyettou A, Medjahed SA. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. International Journal of Computer Applications. 2013;62(1):1-5..

[5] Sathiyanarayanan P, Pavithra S, Saranya MS, Makeswari M. Identification of breast cancer using the decision tree algorithm. In2019 IEEE International conference on system, computation, automation and networking (ICSCAN) 2019 Mar 29 (pp. 1-6). IEEE.

[6] Nemissi M, Salah H, Seridi H. Breast cancer diagnosis using an enhanced extreme learning machine based-neural network. In2018 international conference on signal, image, vision and their applications (SIVA) 2018 Nov 26 (pp. 1-4). IEEE..”

[7] Kaur A, Kaur P. Breast cancer detection and classification using analysis and gene-back proportional neural network algorithm. International Journal of Innovative Technology and Exploring Engineering. 2019 Jun.

[8] Assegie TA. An optimized K-Nearest Neighbor based breast cancer detection. Journal of Robotics and Control (JRC). 2021 May 5;2(3):115-8.

[9] Sharma S, Aggarwal A, Choudhury T. Breast cancer detection using machine learning algorithms. In2018 International conference on computational techniques, electronics and mechanical systems (CTEMS) 2018 Dec 21 (pp. 114-118). IEEE. [10] Y. Ireaneus, A. Rejani, and S. T. Selvi, “EARLY DETECTION OF BREAST CANCER USING SVM CLASSIFIER TECHNIQUE,” 2009.

10.48047/jocaaa.2023.31.04.59

[11] Prasad P. The Application of Machine Learning Techniques to the Diagnosis of Breast Cancer. In2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP) 2023 Mar 18 (pp. 1-7). IEEE..

[12] Nounou MN. Dealing with collinearity in finite impulse response models using Bayesian shrinkage. Industrial & engineering chemistry research. 2006 Jan 4;45(1):292-8.

[13] Berger JO. Statistical decision theory and Bayesian analysis. Springer Science & Business Media; 2013 Mar 14.

[14] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Chapman and Hall/CRC; 1995 Jun 1.

[15] Hopkins B, Geangu E, Linkenauger S, editors. The Cambridge encyclopedia of child development. Cambridge University Press; 2017 Oct 19.

[16] Malluhi B, Basha N, Fezai R, Ibrahim G, Choudhury HA, Challiwala M, Nounou H, Elbashir N, Nounou M. Multiscale bayesian pca for robust process modeling of a fischer–tropsch bench scale process. Chemometrics and Intelligent Laboratory Systems. 2023 Sep 15;240:104921..

[17] Brown PJ, Vannucci M, Fearn T. Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1998;60(3):627-41.

[18] Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018 Jun;12(2):119-26..