EAI Endorsed Transactions
on Pervasive HealthCare and Technology
Research Article EAI.EU

A Comprehensive Feature Engineering Approach for Breast Cancer Dataset

Shambhvi Sharma[1,*] and Monica Sahni[1]

[1]DPS Mathura Road, Delhi, India

Abstract

Breast cancer continues to pose a significant challenge in the field of healthcare, serving as the primary cause of cancer-related deaths in women on a global scale. The present study aims to investigate the intricate relationship between breast cancer, statistical analysis, and feature engineering. By conducting an extensive analysis of a comprehensive dataset and employing sophisticated statistical methodologies, this research endeavor aims to unveil concealed insights that can enrich the medical community's existing knowledge base. Through the implementation of rigorous feature selection and extraction methodologies, the overarching aim is to augment the comprehension of breast cancer. Moreover, the study showcases the successful incorporation of univariate and bivariate analysis in order to enhance the accuracy of diagnostic procedures. The convergence of these disciplines exhibits considerable promise in the realm of breast cancer detection and prediction, facilitating cooperative endeavours aimed at addressing this widespread malignancy.

Keywords: Breast Cancer, Univariate Analysis, Bivariate Analysis, Heat Map, Correlation

Received on 08 December 2023, accepted on 29 February 2024, published on 07 March 2024

doi: 10.4108/ eetpht.10.5327

1. Introduction

Breast cancer is a significant challenge in the field of healthcare, and one of the prominent reasons of cancer-related casualties among women worldwide. The statistics presented serve as a poignant reminder of the relentless impact of breast cancer. According to available data, in 2023, it is estimated that 352,510 women in the United States only are diagnosed with breast cancer [1]. The pressing need to thoroughly understand this medical condition, along with the necessity to develop more accurate methods for diagnosis and treatment, has driven the investigation of advanced analytical techniques in the field of medical research.

The field of cancer research has seen a significant increase in the importance of statistical analysis, which plays a fundamental role in the investigation and understanding of this disease [2]. The comprehensive examination of various elements that contribute to the initiation and advancement of breast cancer calls for a methodical inquiry that surpasses the conventional confines of medical research. Through the utilization of statistical methodologies, both researchers and healthcare practitioners possess the capacity to delve into the complex intricacies of the disease. This allows for the meticulous examination of patterns, the identification of risk factors, and ultimately the improvement of prognosis accuracy [3].

The identification and extraction of significant features from complex datasets represents a critical milestone in the field of statistical analysis. The process described, which holds significant importance, is deeply embedded in the field of feature engi-neering. It empowers researchers by equipping them with the necessary means to extract the most relevant characteristics from extensive and frequently complex datasets [4]. In the realm of breast cancer analysis, this undertaking ultimately results in the identification of factors

*Corresponding author. Email: shambhvi.sharma@gmail.com

that possess predictive capabilities, facilitating well-informed decision-making within clinical environments. By combining advanced algorithms with carefully curated datasets, researchers have discovered promising opportunities to detect breast cancer at its early stages [5]. This breakthrough has the potential to enhance the effectiveness of treatment interventions significantly.

The present research paper endeavors to explore the intersection of breast cancer, statistical analysis, and feature engineering. Through the utilization of a comprehensive dataset about breast cancer, our objective is to elucidate the latent insights that can be derived from conventional statistical methodologies, thereby contributing to the knowledge base of the medical community. By employing a rigorous methodology involving meticulous feature selection and extraction techniques, the primary objective of our investigation is to make a valuable contribution to the existing body of scientific knowledge of breast cancer. The integration of these various fields of study shows great potential in transforming the field of breast cancer diagnosis and prognosis, highlighting the collaborative efforts aimed at reducing the impact of this widespread malignancy.

The manuscript is divided into the following sections. After the detailed introduction of breast cancer, applicability of machine learning, next section demonstrates the related work, proposed methodology, results, and discussion. Finally, the conclusion and future scope is provided.

2. Related Work

The section presents a comprehensive summary of the research done on breast cancer data to explore the possibility of statistical analysis of breast cancer data. The recognition of the necessity to utilize statistical methodologies to uncover the inherent patterns and influential factors is readily apparent in the extensive body of research that has been produced throughout the years.

The importance of statistical analysis within the realm of breast cancer research has been emphasized by numerous studies that have aimed to unravel the complex nature of this ailment. Previous studies have underscored the sombre actuality of breast cancer's position as the most formidable opponent in the realm of women's health, exhibiting a widespread prevalence on a global scale. Bray et. al. (2018) conducted a study on a specific topic. The global cancer statistics for the year 2018 were obtained from GLOBOCAN, an organization that specializes in estimating the incidence and mortality rates of various types of cancer on a global scale. These estimates encompassed data from 185 countries and covered a total of 36 different types of cancer [6].

The significance of statistical analysis in unravelling the intricate interplay involved in the development of breast cancer has been effectively showcased in previous investigations, such as those carried out by Chen et al. (2016) [7] and Anderson et al. (2019) [3]. Chen et al. conducted an in-depth investigation into the complex realm of gene expression profiles to ascertain significant biomarkers linked to different subtypes of breast cancer. To achieve this, they utilized sophisticated statistical methodologies including principal component analysis (PCA) and hierarchical clustering [2]. In a study conducted by Anderson et al., the researchers aimed to investigate the influence of demographic and clinical variables on breast cancer survival rates. To achieve this, they employed Cox proportional hazards models, a statistical method commonly used in survival analysis [8]. The studies collectively underscore the intrinsic potency of statistical analysis in elucidating the complexities associated with the fundamental mechanisms underlying breast cancer.

Furthermore, the interdependent association between statistical analysis and ma-chine learning methodologies has given rise to a novel paradigm in the field of breast cancer research. The integration of these fields, as demonstrated in the studies con-ducted by Li et al. (2018) [9] and Wang et al. (2020) [10], has given rise to an innova-tive methodology for the identification and forecasting of diagnoses and prognoses. In their study, Li et al. employed ensemble learning techniques as a means to aug-ment the discriminatory capabilities of classification models. In a similar vein, Wang et al. undertook an investigation aimed at elucidating the capabilities of deep learning architectures in the realm of breast cancer histopathology image analysis. Their study highlighted the considerable promise of deep convolutional neural networks (CNNs) in the realm of automating tumor identification [11].

In the realm of feature selection and extraction, the investigation conducted by Zhu et al. (2017) [6] has shed light on the profound impact of this procedure when integrated with machine learning techniques. The study conducted by Zhu et al. (year) demonstrated the effectiveness of employing recursive feature elimination (RFE) in conjunction with support vector machines (SVMs) to identify the most optimal subsets of features in breast cancer classification. This research significantly contributes to the improvement of diagnostic accuracy in this domain [6].

The studies presented collectively emphasize the diverse range of applications for statistical analysis within the field of breast cancer research. The integration of statistical techniques with machine learning methodologies signifies a significant advancement in the field. This development holds great promise for the early identification, accurate diagnosis, and focused treatment of breast cancer. Ultimately, this collaborative effort aims to mitigate the substantial impact of this disease on women’s health.

3. Proposed Methodology

In this section, the methodology to be employed in our research is outlined, to comprehensively analyse breast cancer data. The comprehensive research framework incorporates various methodologies such as statistical analysis, feature selection, and feature extraction techniques. The primary objective of our research is to demonstrate the practicality of employing these methodologies to improve the comprehension and diagnostic capabilities of breast cancer

datasets. Figure 1 demonstrates the steps employed in the methodology of the research work.

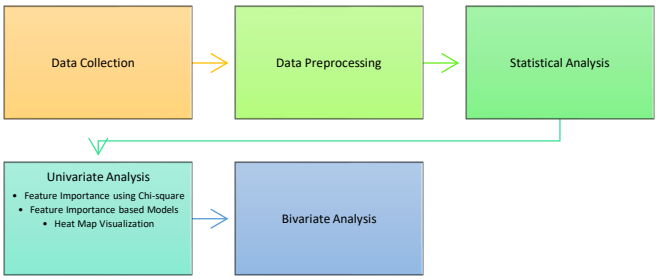

Figure 1. Steps demonstrating the proposed Methodology

The methodology is initiated by conducting a comprehensive statistical analysis of the breast cancer dataset. The primary objective of this phase is to extract significant insights and identify prevailing patterns from the dataset by utilizing essential statistical methodologies. The utilization of descriptive statistics is imperative to obtain a comprehensive understanding of the dataset. This entails the calculation of various measures, including means, medians, and standard deviations. The forthcoming investigation will be complemented by the utilization of data visualization methodologies, such as histograms, box plots, and scatter plots, to illuminate the dispersion patterns of various attributes and ascertain any potential associations between them. The steps are further elaborated as under:

1. **Data Collection:** The first step is to collect the data. The data collected is the famous Breast Cancer Wisconsin (Diagnostic) Data Set from Kaggle [12]. The collected dataset has 569 rows and 33 columns. The independent features are the features related to breast cancer whereas the dependent feature is the diagnosis features which consist of two classes namely: benign and malignant. Figure 2 represents the distribution of the classes in the dataset.
2. **Data Preprocessing:** The second step is to preprocess the data. This involves handling null values, checking outliers, and checking categorical values. Since the dataset does not contain any categorical values and outliers the only step performed is to handle the null values [13]. All the null values were dropped.
3. **Statistical Analysis:** The methodology is initiated by conducting a comprehensive statistical analysis of the breast cancer dataset. The primary objective of this phase is to extract significant insights and identify prevailing patterns from the dataset by utilizing essential statistical methodologies. The utilization of descriptive statistics is imperative in order to obtain a comprehensive understanding of the dataset. This entails the calculation of various measures, including means, medians, and standard deviations [14]. The forthcoming investigation will be complemented by the utilization of data visualization methodologies, such as histograms, box plots, and scatter plots, in order to illuminate the dispersion patterns of various attributes and ascertain any potential associations between them.

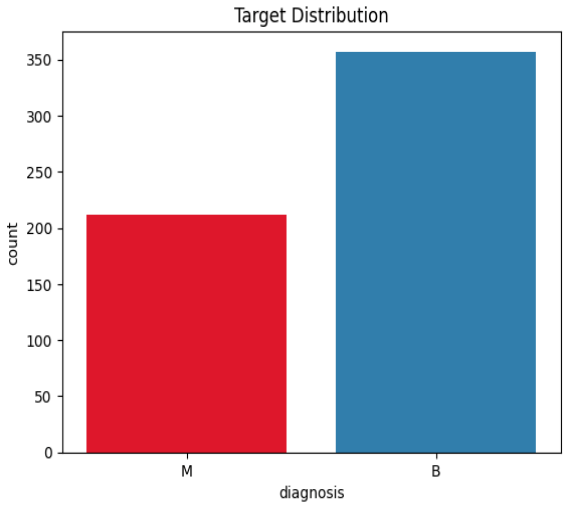

Figure 2. Class Distribution of the Breast Cancer Dataset

4. **Univariate Analysis:** These techniques are employed after the preliminary statistical analysis in order to improve the overall quality of the dataset. The objective of these techniques is to ascertain the most pertinent and distinguishing characteristics that substantially contribute to the categorization of instances of breast cancer [15]. In this study, various methodologies will be examined, including:

- Correlation analysis based on chi-square test. The correlation is evaluated on the basis of the p-value. If the p-value is higher than 0.5, then it is considered to be a positive correlation, otherwise the correlation between the factors is ignored.

- Correlation based on the tree-based models. Tree based models are employed to find out the correlation between the independent factors and the dependent factors.

- Visualization with the Heat Map: A heat map is visualized based on the color palette to show the correlation between the factors.

5. **Bivariate Analysis**: The next step is to perform the analysis of two independent factors at the same time. This approach enables us to effectively encapsulate the inherent structure of the dataset by projecting it onto a lower-dimensional space characterized by

orthogonal axes. The process of reducing the dimensionality of the dataset serves to decrease its complexity while preserving its essential characteristics.

In summary, the methodology we propose aims to effectively combine statistical analysis, feature selection, and feature extraction techniques in order to uncover the underlying patterns present in breast cancer data. Through the demonstration of the practicality and effectiveness of these methodologies, our objective is to make a valuable contribution to the existing repertoire of tools accessible to medical researchers and practitioners. This contribution will serve to further advance the field of breast cancer diagnosis and treatment strategies.

4. Results and Discussion

The current section demonstrates the results of applying the various methodology steps on the chosen dataset and the discussion of the results is provided thereafter. Statistical tests are a valuable tool for identifying and selecting features that exhibit a robust relationship with the output variable. By applying these tests, researchers can discern which features demonstrate the strongest association with the outcome of interest. This process aids in the identification of key variables that contribute significantly to the predictive power of a model or the understanding of a phenomenon.

4.1 Univariate Analysis

Univariate selection is a widely utilized statistical technique in research that aims to identify the most pertinent features or variables that exhibit significant contributions to a specific outcome. The SelectKBest class, which is part of the scikit-learn library, provides researchers with a variety of statistical tests that can be utilized to choose a specific number of features from a given dataset. The notion of feature importance pertains to the assessment and quantification of the relative significance or contribution of individual features or variables within a specific dataset. The acquisition of feature importance for each feature in a given dataset can be achieved by utilizing the feature importance property of the model. The notion of feature importance involves the allocation of a numerical score to each feature present in a given dataset. The score provided in this context functions as a metric that signifies the relative significance or pertinence of a specific feature with the target variable of interest. In the context of this study, it is important to note that a higher score is indicative of a heightened level of significance or pertinence attributed to the particular feature under consideration.

One additional approach utilized for determining the significance of each feature within a dataset involves leveraging the feature importance attribute of the model. The concept of feature importance entails assigning a score to each feature within a dataset. A higher score indicates a

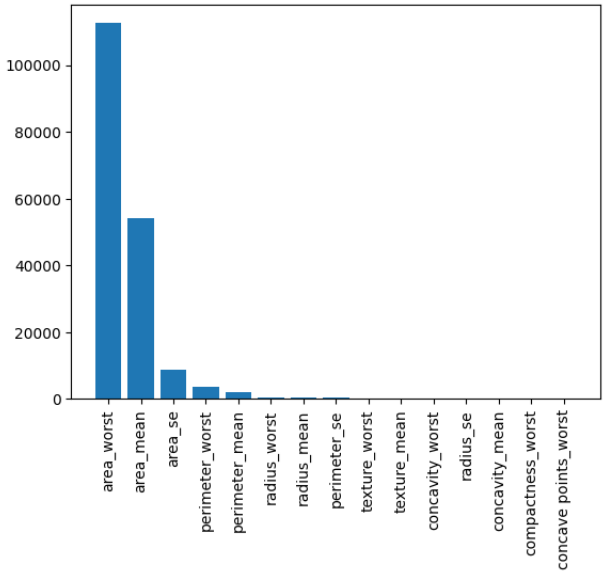

Figure 3. Top 15 Features Extracted based on Chi-Square Test

greater degree of importance or relevance of the feature concerning the output variable. The feature importance attribute is a pre-existing class that is included in Tree Based Classifiers. In this study, the authors have utilized the Extra Tree Classifier to extract the top 10 features.

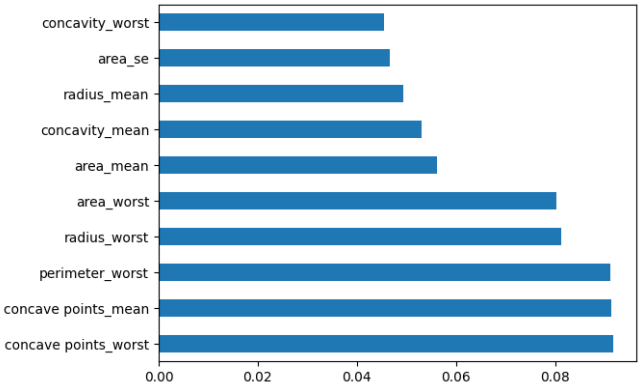

Figure 4. Top 10 features extracted based on Tree-Based Models

The correlation matrix, accompanied by a heatmap, is a visual representation of the pairwise correlations between variables in a dataset. It provides valuable insights into the strength and direction of relationships between variables, aiding in the identification of patterns and dependencies.

Correlation is a statistical measure that quantifies the relationship between features or the target variable. The relationship between two variables can exhibit either a positive or negative correlation. A positive correlation is observed when there is a tendency for the value of one variable to increase as the value of another variable increases.

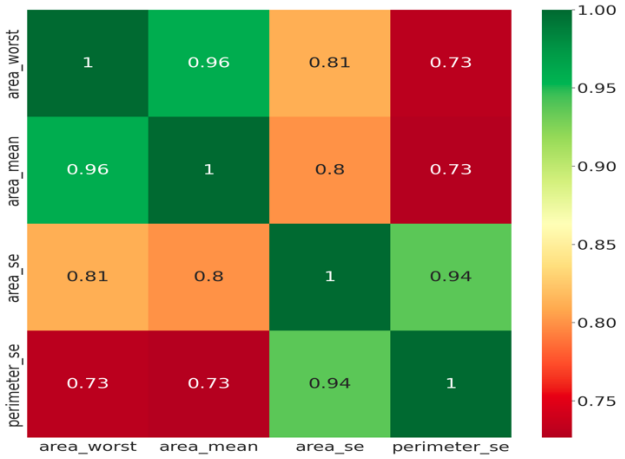

Figure 5. Heat-Map Visualization based on Top 4 Features

In contrast, a negative correlation is indicative of an inverse relationship, wherein an increase in the value of one variable is typically accompanied by a decrease in the value of the other variable. The application of a heatmap enables the analysis of the characteristics that display the greatest level of correlation with the target variable.

4.2 Bivariate Analysis

Various feature selection techniques have been developed and widely used in the field of machine learning. These techniques aim to identify and select a subset of relevant features from a given dataset, while discarding irrelevant or redundant ones. The selection of appropriate feature selection techniques is crucial, as it directly impacts the performance and interpretability of the resulting models. In the pursuit of simplicity and effectiveness, researchers have developed several feature selection techniques that are known for their ease of use and ability to yield satisfactory results. These techniques typically employ different strategies to evaluate the relevance and importance of individual features, and subsequently rank.

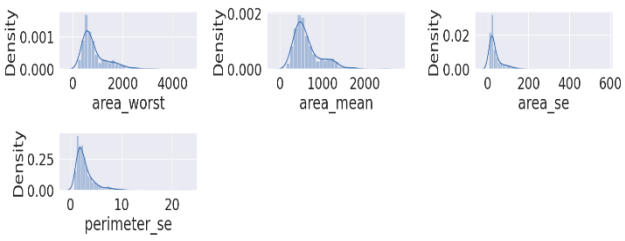

Figure 6. Heat-Map Visualization based on Top 4 Features

6. Conclusion

The research elucidated in this scholarly article highlights the fundamental significance of statistical analysis and feature engineering in propelling the progress of breast cancer research and diagnosis. By employing a meticulous approach encompassing extensive analysis of a comprehensive dataset, utilization of univariate and bivariate techniques, and integration of machine learning methodologies, this investigation has yielded valuable insights into the intricate dynamics of the underlying mechanisms associated with breast cancer. The discovery of influential features and relationships has provided insights into potential avenues for early detection and improved prognostic accuracy. Through the integration of statistical analysis, feature engineering, and machine learning, this research endeavor establishes a foundation for enhancing medical decision-making processes and facilitating the development of personalized treatment strategies. The persistent prevalence of breast cancer among women necessitates a concerted and collaborative endeavor aimed at mitigating its impact on both global healthcare systems and individuals. The future of breast cancer research holds immense promise, as it offers numerous opportunities to enhance existing methodologies, broaden datasets, and investigate innovative techniques. These endeavors will undoubtedly contribute to the ongoing advancement of our knowledge and comprehension of this complex disease.

References

[1] N. Sharma, M. Mangla, M. Ishaque and S. N. Mohanty, "Inferential Statistics and Visualization Techniques for Aspect Analysis," 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 2023, pp.

[2] https://www.cancer.net/cancer-types/breast-cancer/statistics.

[3] Dubey, A. K., Gupta, U., & Jain, S. (2015). Breast cancer statistics and prediction methodology: a systematic review and analysis. Asian Pacific journal of cancer prevention, 16(10), 4237-4245.

[4] Lewis, J. T., Hartmann, L. C., Vierkant, R. A., Maloney, S. D., Pankratz, V. S., Allers, T. M., ... & Visscher, D. W. (2006). An analysis of breast cancer risk in women with single, multiple, and atypical papilloma. The American journal of surgical pathology, 30(6), 665-672.

[5] YK Ng, LN Ung, FC Ng, LSJ Sim, E. (2001). Statistical analysis of healthy and malignant breast thermography. Journal of medical engineering & technology, 25(6), 253-263.

[6] Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2(2011), 37-45.

[7] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394-424.

[8] Chen, H., Boutros, P. C., & Vennettilli, A. (2016). Characterizing heterogeneous subtype by integrating gene expression data and pathway markers. BMC Bioinformatics, 17(Suppl 13), 323.
[9] Anderson, W. F., Luo, S., Chatterjee, N., Rosenberg, P. S., & Matsuno, R. K. (2019). J Natl Cancer Inst, 111(3), 310-320.
[10] Li, H., Pang, B., & Wu, N. (2018). A hybrid method for breast cancer diagnosis based on feature selection and ensemble learning. Frontiers in Genetics, 9, 597.
[11] Wang, X., Janowczyk, A., Zhou, Y., Thawani, R., Fu, P., Schalper, K., ... & Yao, J. (2020). Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. Scientific Reports, 10(1), 1-12.
[12] Zhu, W., Zeng, N., Wang, N., Yang, Y., & Wu, F. (2017). A review on region-based object detection algorithms. Pattern Recognition, 70, 167-183.
[13] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[14] D. Jalan, A. Tuli, V. Chaudhary, N. Sharma and M. Rakhra, "Machine Learning Models for Life Expectancy," 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/ICAIA57370.2023.10169737.