# Breast Cancer Classification Using Ensemble Voting: A Feature Selection Approach

Antu Kumar Guha[1], Jun-Jiat Tiang[2]*, Abdullah-Al Nahid[3]*
Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208, Khulna, Bangladesh[1,3]
Centre for Wireless Technology-CoE for Intelligent Network-Faculty of Artificial Intelligence and Engineering,
Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia[2]

***Abstract*—Breast cancer is one of the most common and deadly diseases affecting women around the worldwide. It is specially affecting in regions where has limited access to advanced diagnostic tools. Recent studies have shown that blood-based biomarkers can give a cost-effective alternative for early detection. This paper represents a machine learning-based approach for classifying breast cancer using clinical and biomedicial data. We have used the Breast Cancer Coimbra dataset for our study. We employed four filter-based feature selection methods—Mutual Information, Chi-Square, ANOVA F-test, and Pearson Correlation Coefficient—to identify the most relevant features for classification. We have applied two classifiers (AdaBoost and Ensemble Voting Classifier) to enhance predictive accuracy. The ensemble model achieved an accuracy of 82.86%. Key features such as glucose, HOMA, insulin, resistin, and age consistently contributed across all selected methods.It highlights that a few of the features has a great significance in breast cancer prediction. This study also try to investigate the reasons behind the missclassification cases. Our results show that using statistical feature selection with ensemble learning reasonable helps to boost the accuracy of breast cancer prediction. This approach helps the model focus on the most important features.**

***Keywords*—Breast cancer; machine learning; feature selection; ensemble learning; AdaBoost; biomedical data classification**

## I. Introduction

Cancer is a deadly disease in which some of the body's cells grow uncontrollably and spread to other parts of the body [1]. According to World Health Organization (WHO), 10 million people died of cancer in 2020 [2]. It is nearly a sixth of all cancer deaths. There are over 100 different kinds of cancer. One of these cancers is breast cancer (BC) which is mainly affecting women.

In 2022, the world cancer research fund estimated that there were nearly 2,296,840 new BC cases [3].It was nearly 11.6 percent of all cancer cases. Moreover, almost 670 000 deaths were reported because of BC in the same year by WHO [4]. Fig. 1 illustrates the total incidence of breast cancer cases worldwide between the years 2015 and 2024. It has seen more cases over the years. In 2015, there were close to two million cases. It grew slowly every year, until it hit 2.5 million in 2024. This number remained the same between 2020 to 2022, potentially attributed to the impact of the COVID-19 pandemic on health services [5]. Then the numbers began to climb again. This increase explains the need for early detection and better awareness in order to reduce the risk of breast cancer [6]. Metastasis and late detection is one of the major factors that leads to high mortality among BC patients. Metastasis refers to the motion of cancer in the breast to other body parts of a human body. The metastasis increases the death rate. The discovery of BC at its early stage could save lives and make treatment effective. We can use various screening methods such as mammography, magnetic resonance imaging (MRI), ultrasound, and biopsy for BC detection [7].

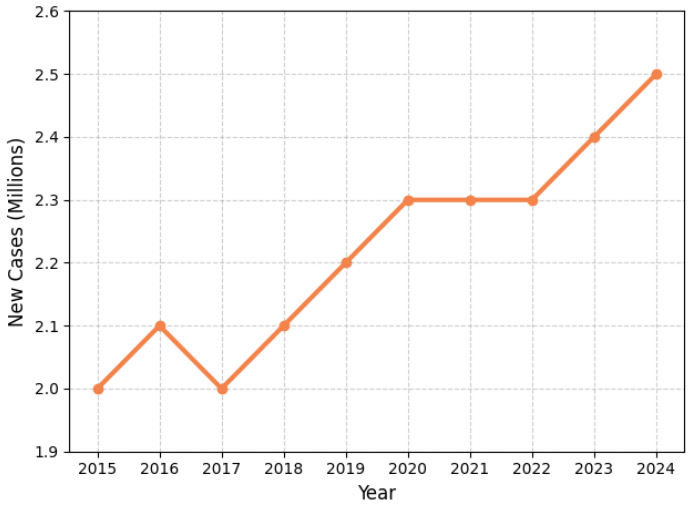


Fig. 1. Number of BC cases globally for last 10 years.

However, these methods are too expensive to afford for many women in rural area.

Also, villages lack professional physicians and adequate medical equipment [8]. The rural area mortality rate is therefore likely to be 2-4 percent higher than the urban regions. In order to address this challenge, scientists have looked at other cheaper diagnostic methods. Blood tests for specific biomarkers—like glucose, insulin, leptin, adiponectin, and others—do not directly diagnose breastcancer. But, these biomarkers can provide indicative information of metabolic health as well as inflammation within the body. This information allows to assess individual risk of disease development or rate of progression over time. The identification of these biomarkers may improve access to care as the blood tests are easy to perform in rural environments.

The Breast Cancer Coimbra (BCC) dataset is a popular dataset for the BC classification problem. This dataset consists of nine features (seven blood-based markers, age, and BMI). Glucose level is a measure of sugar in the human blood stream.

*Corresponding authors.

All cells in the body use glucose for energy. Hormones such as insulin and glucagon control glucose levels in our body. As a glucose test is also critical to the detection and monitoring of breast cancer. Unlike in normal cells, cancer cells metabolize glucose at a more rapid rate. Increased glucose utilization can be detected by means of glucose tests. Leptin is a hormone produced by adipose tissue that is involved in the regulation of appetite, energy expenditure, and metabolism more generally [9], [10]. Leptin can play a significant role in breast cancer detection and progression.

Resistin is a hormone which is mainly released immune cells in humans. It causes inflammation and helps cancer cells grow and spread [11]. High resistin levels are often found in people with cancer. Adiponectin is a hormone produced by fat cells. It is also very important to control blood sugar and reduce inflammation. Insulin is a hormone produced by the pancreas for regulating blood sugar. It allows the body to burn sugar for energy or store it for use later. Insulin also is postulated to contribute to breast cancer development at high levels because it promotes rapid growth of cancer cells [12], [13]. Checking insulin levels may help with early detection and prevention. The Homeostatic Model Assessment (HOMA) is a way to measure how resistant the body is to insulin (HOMA-IR) and how well the insulin-producing cells (beta-cells) are working (HOMA-B) [14]. Because insulin resistance and high blood sugar are linked to a higher risk of breast cancer, HOMA can help in early detection and assessing risk [14]. Monocyte Chemoattractant Protein-1 (or MCP-1) is a molecule that is very important in inflammation and immune system [15]. More recent studies show that MCP-1 is also involved in detecting breast that cancer, disease progression, and its spread (metastasis) [15].

Machine learning (ML) has shown promising performance in the prediction of early stage breast cancer in recent years, using blood based biomarkers. By using the Breast Cancer Coimbra Dataset (BCCD) Hernández-Julio et al. reached an accuracy of 95.90% using a novel approach of clustering combined with the use of pivot tables, the results were assessed through 10 folds cross validation [16]. Singh also used k-NN and achieved a remarkable accuracy of 92.11% with a 67-33 training-testing split. His research results showed to contribute to improve prediction including the use of specific features such as BMI and resistin [17]. Polat and Senturk developed a new hybrid model based on Median Absolute Deviation, feature weighting through K-means clustering and AdaBoost classifier. This method achieved 91.37% of accuracy [18]. Likewise, Akben constructed a rule based expert system implementing decision trees that achieved an accuracy of 90.52% and found important ranges of values for hormone and obesity-related variables that are linked to breast cancer [19]. Islam and Poly also utilized the k-NN algorithm with 10 folds of cross-validation and obtained an accuracy of 86.00% [20].

While earlier work achieved strong accuracy through individual classifiers or ensemble techniques of hybrid normalization. Among them few have successfully merged various statistical feature selection methods to enhance model robustness. In addition, there was low focus on applying these feature selection methodologies to an ensemble setting. To fill this gap, the present work suggests an ensemble voting classifier with four filter-based feature selection algorithms-Mutual Information, Chi-Square, ANOVA F-test, and Pearson Correlation Coefficient-added to the AdaBoost classifier. This ensures that the most significant biomarkers are used in training, improving accuracy of predictions and interpretability.

This study contributes in three main ways.

*1) Methodological:* We build a new AdaBoost-based ensemble that modifies a range of different filter feature-selection methods instead of changing only the classifiers. This makes the model more precise and stable.

*2) Analytical:* By comparing the output of different feature selectors, we find a group of major biomarkers (*Glucose, HOMA, Insulin, Resistin, and Age*) that appear repeatedly and are biologically relevant to breast cancer.

*3) Applicational:* Since our model uses simple filtering methods, it runs very fast and can be used even in clinics or hospitals with limited computing facilities.

Overall, this study clearly explains the new ideas and practical benefits of our approach from both the artificial intelligence and medical points of view.

Moreover, the proposed framework identifies significant advantages such as efficient computation, understandable model behavior for interpretability, and stable performance with limited data. With many light-weight statistical feature-selection methods integrated via AdaBoost, the model achieves fast training and inference without compromising interpretability. The ensemble structure also ensures stability and generalization when there are only a few medical samples. These techniques increase the uniqueness of this study and prove its utility for real-world breast cancer prediction applications.

## II. Methodology

This study has used machine learning techniques to classify breast cancer cases based on clinical and biochemical data. Fig. 2 shows the overall workflow of the proposed breast cancer classification model. The process includes data preprocessing, feature selection using four statistical methods (Mutual Information, Chi-Square, ANOVA, and Correlation), AdaBoost-based classification, and ensemble voting to produce the final prediction. AdaBoost is applied due to its strong performance in binary classification problems. The feature selection techniques help identify which features are most relevant before training the models.

### A. Dataset

We used the publicly available Breast Cancer Coimbra dataset from the UCI Machine Learning Repository, collected at the University Hospital Center of Coimbra (Portugal).This dataset is selected because it provides clinically interpretable, low-cost, blood-based biomarkers in support of our aim of constructing effective and affordable cancer-screening models. This dataset contains 116 instances. Among the 116 participants, 64 were diagnosed with breast cancer, and 52 were not. Each instance in the dataset includes nine features: Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP.1, along with a classification label. If the BC is presence the classification lavel is 2 otherwise 1. The features are all numerical and take from routine blood analysis and

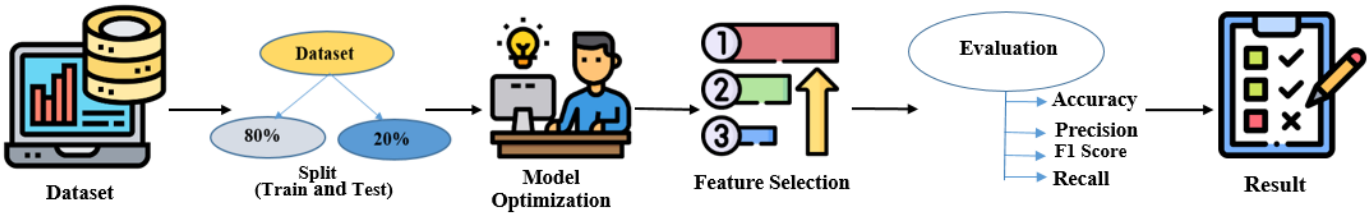


Fig. 2. The process of detecting BC.

TABLE I. FEATURES OF THE BREAST CANCER COIMBRA (BCC) DATASET

| Feature | Description | Lowest Value | Highest Value | Unit |
|---|---|---|---|---|
| Age($f_1$) | Age of the patient | 24 | 89 | Years |
| BMI($f_2$) | Body Mass Index | 18.37 | 37.10 | kg/m$^2$ |
| Glucose($f_3$) | Blood glucose level | 70 | 201 | kg/dL |
| Insulin($f_4$) | Blood insulin level | 2.43 | 58.46 | $\mu$U/mL |
| HOMA($f_5$) | Homeostatic Model Assessment | 0.50 | 25.05 | (dimensionless) |
| Leptin($f_6$) | Leptin hormone level | 6.33 | 90.28 | $\mu$g/mL |
| Adiponectin($f_7$) | Adiponectin hormone level | 1.65 | 33.75 | $\mu$g/mL |
| Resistin($f_8$) | Resistin hormone level | 3.21 | 55.21 | ng/mL |
| MCP-1($f_9$) | Monocyte Chemoattractant Protein-1 | 90.09 | 1698.44 | pg/dL |

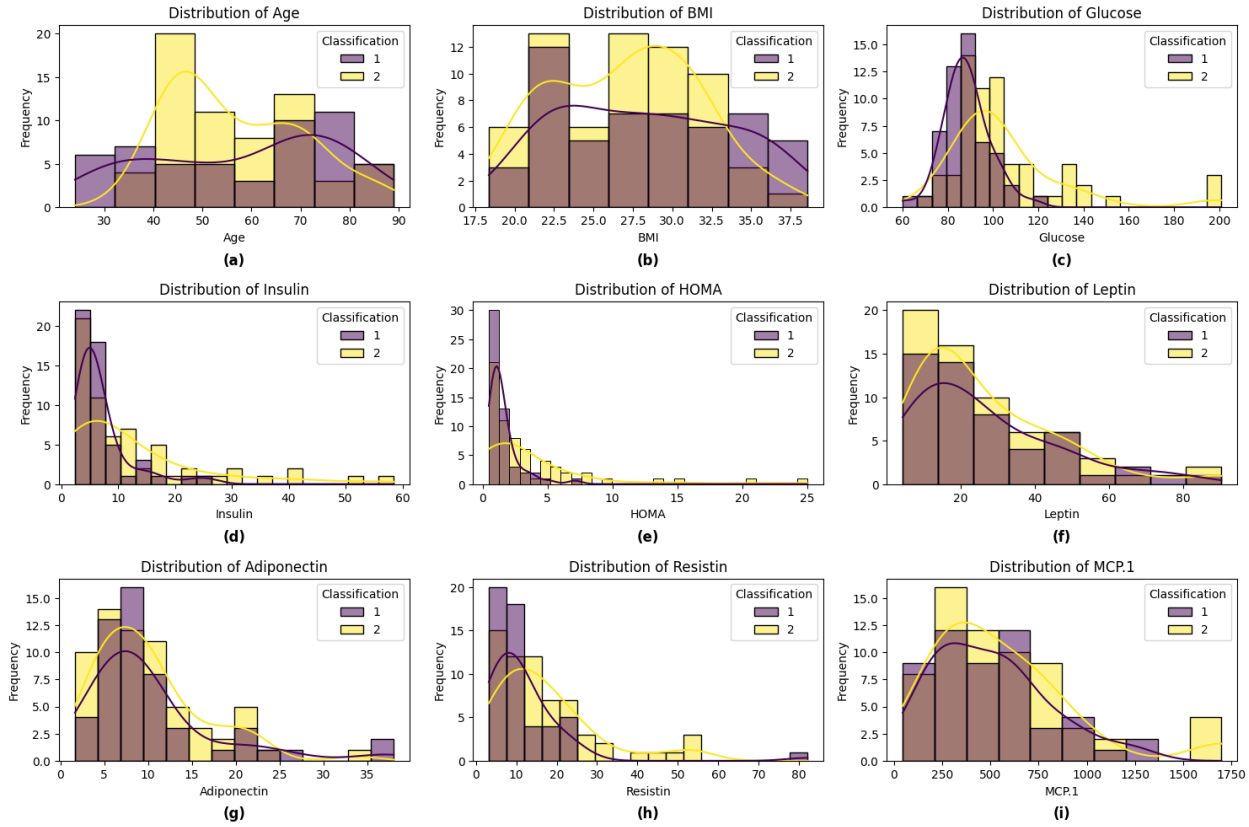


Fig. 3. Distribution of the features.

clinical measurements. According to Table I we get the highest value,lowest value and units of the features. Here,The average age of participants is about 57 years, with a minimum of 24 and a maximum of 89. BMI ranges from around 18.3 to 38.6, and glucose levels vary significantly—from 60 to 201 mg/dL, with an average of nearly 98 mg/dL. Insulin levels also show a wide range, from just 2.43 to over 58 µU/mL. Other features like HOMA, Leptin, and MCP.1 also display

noticeable variation, reflecting differences in metabolic and hormonal profiles among individuals.

Fig. 3 displays the distribution of all nine clinical and biochemical features of the Breast Cancer Coimbra (BCC) dataset, plotted separately for each class (Class 1 = no breast cancer, Class 2 = breast cancer). These histograms with kernel density estimates show the variability and overlap between features. As seen in Fig. 3 some biomarkers such as Glucose, HOMA, and Resistin exhibit evident differences between the two classes, disclosing their high potential as discriminative predictors. On the other hand, features such as Leptin and Adiponectin show significant overlap, which suggests less separability. Overall, the figure shows how statistical feature selection methods can strengthen attributes with distinctive inter-class variation, and thereby increase model accuracy and interpretability.

Before training the models, we have split the dataset into training and testing sets to evaluate the model performance. We have used various metrics such as accuracy, precision, recall, and F1-score to measure how well each classifier performs on unseen data.

### B. Feature Selection

Feature selection means picking the most useful features from a dataset to help a model perform better [21].In this study, we used four common filter-based feature selection methods on the Breast Cancer Coimbra dataset. The goal was to improve the model's accuracy and remove features that are not helpful or needed.By doing this, we aim to find the most important features that help predict if a patient has breast cancer or not [22].

*1) Mutual Information (MI):* Mutual Information measures how much information about the output (breast cancer status) is shared with each input feature. In simple terms, it tells us how strongly a feature is related to the target. A higher MI score means the feature has a stronger dependency with the target and is more useful for prediction. We selected features with the highest MI scores.

*2) Chi-square test:* The Chi-Square test checks whether there is a significant relationship between a feature and the target. It works well with categorical data. Features that show a higher Chi-Square score are considered to have a stronger relationship with the target and therefore are selected.

*3) ANOVA (Analysis of variance):* Analysis of Variance (ANOVA) is a popular statistical method that checks if the average (mean) value of a feature is very different between two or more groups or classes [23].

Here the test statistic F-value is defined as:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

A high F-value indicates that the feature's mean differs notably between classes. The associated p-value is used for decision-making:

- A small p-value ($p < 0.05$) suggests the feature is important for distinguishing between classes.
- A large p-value implies the feature is not useful.

In feature selection, features with low p-values and high F-values are preferred.

### C. Correlation Coefficient Method

The Correlation Coefficient method measures the linear relationship between a feature and the target variable. It is usually computed using the Pearson correlation coefficient [24]:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{1}$$

where $x_i$ and $y_i$ are individual values, and $\bar{x}$, $\bar{y}$ are their means. A higher absolute value of $r$ indicates a stronger relationship. Features with high correlation (positive or negative) to the target are selected [25].

## III. Results

For selecting the relevant features for this work we tested four statistical filtering techniques: Mutual Information, Chi-Squared Test, F-Test(ANOVA) and Correlation. All three approaches appraise feature importance from a unique statistical standpoint. Mutual Information measures how much information a feature provides about the output. From Fig. 4 according to mutual information method, Age is the most informative feature. It has the highest mutual information score. That means it may be most useful for prediction. Other features like HOMA, Resistin, and Glucose also show some importance. However, features such as BMI, Insulin, and Adiponectin have scores of zero, meaning they don't help much in this case. The chi-square test checks if there is a relationship between each characteristic and breast cancer. Similarly, from Fig. 4 we also see that according to Chi-Square test Insulin and MCP.1 have the highest scores with very low p-values, which means they are strongly related to BC. Glucose, HOMA, and Resistin are also important. On the other hand, Age and Adiponectin are not very useful according to this test. The F-Test compares the average values of features between groups. It showed that Glucose, HOMA, and Insulin are the top features. Features like Adiponectin and MCP.1 have very small F-scores. It means these features didn't vary much and not very helpful in BC detection. According to Correlation Coefficient feature selection Glucose, HOMA, and Insulin again come out on top, showing a moderate relationship with the disease. Features like Leptin and Adiponectin had values near zero. So these are not useful for prediction. Here, we have used a stepwise feature addition approach to evaluate model performance across various combinations of features. The features are added one by one according to the MI score, Chi-2 score, ANOVA f-value and correlation score. Fig. 5 shows how the model performance changes as features are added stepwise based on different selection methods. The graph clearly illustrates that adding features such as Resistin and Insulin notably improves accuracy, while including less relevant biomarkers causes a slight decline.

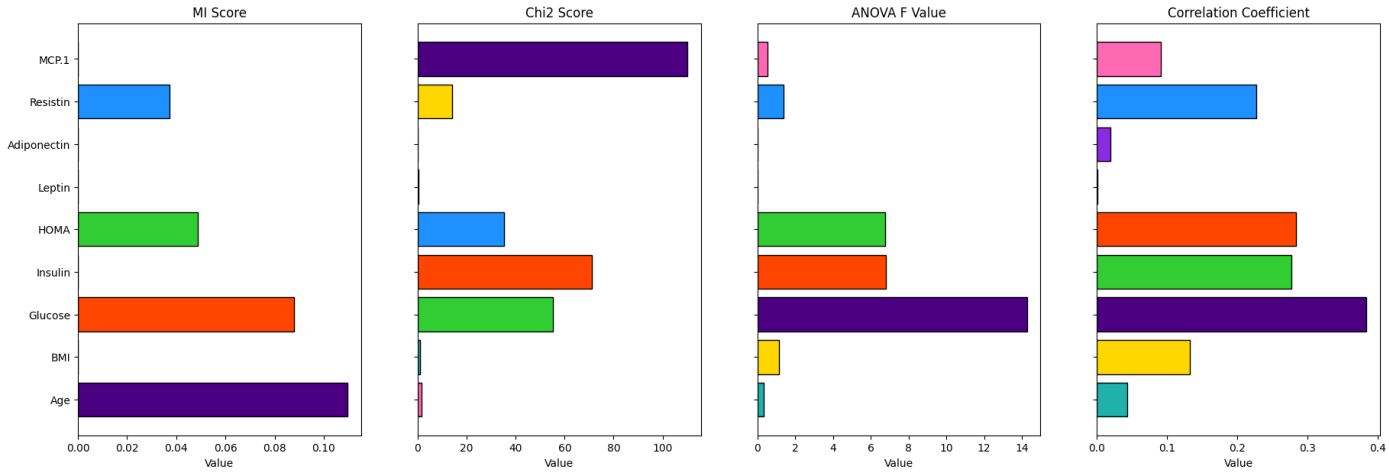


Fig. 4. Comparison of feature importance using MI score, Chi² value, ANOVA F-Statistic, and correlation coefficient.

TABLE II. FEATURE SELECTION ACCORDING TO MUTUAL INFORMATION (MI) SCORE, CHI-SQUARE SCORE, ANOVA F-VALUE, AND CORRELATION COEFFICIENT

| i | **Mutual Info** ($C_{\text{mi}i}$) | **Chi-2** ($C_{\chi i}$) | **ANOVA F-value** ($C_{\text{f-value}i}$) | **Correlation** ($C_{\text{correlation}i}$) |
|---|---|---|---|---|
| 1 | $\{f_1\}$ | $\{f_9\}$ | $\{f_3\}$ | $\{f_3\}$ |
| 2 | $\{f_1, f_3\}$ | $\{f_9, f_4\}$ | $\{f_3, f_4\}$ | $\{f_3, f_5\}$ |
| 3 | $\{f_1, f_3, f_5\}$ | $\{f_9, f_4, f_3\}$ | $\{f_3, f_4, f_5\}$ | $\{f_3, f_5, f_4\}$ |
| 4 | $\{f_1, f_3, f_5, f_8\}$ | $\{f_9, f_4, f_3, f_5\}$ | $\{f_3, f_4, f_5, f_8\}$ | $\{f_3, f_5, f_4, f_8\}$ |
| 5 | $\{f_1, f_3, f_5, f_8, f_2\}$ | $\{f_9, f_4, f_3, f_5, f_8\}$ | $\{f_3, f_4, f_5, f_8, f_2\}$ | $\{f_3, f_5, f_4, f_8, f_2\}$ |
| 6 | $\{f_1, f_3, f_5, f_8, f_2, f_4\}$ | $\{f_9, f_4, f_3, f_5, f_8, f_1\}$ | $\{f_3, f_4, f_5, f_8, f_2, f_9\}$ | $\{f_3, f_5, f_4, f_8, f_2, f_9\}$ |
| 7 | $\{f_1, f_3, f_5, f_8, f_2, f_4, f_6\}$ | $\{f_9, f_4, f_3, f_5, f_8, f_1, f_2\}$ | $\{f_3, f_4, f_5, f_8, f_2, f_9, f_1\}$ | $\{f_3, f_5, f_4, f_8, f_2, f_9, f_1\}$ |
| 8 | $\{f_1, f_3, f_5, f_8, f_2, f_4, f_6, f_7\}$ | $\{f_9, f_4, f_3, f_5, f_8, f_1, f_2, f_6\}$ | $\{f_3, f_4, f_5, f_8, f_2, f_9, f_1, f_7\}$ | $\{f_3, f_5, f_4, f_8, f_2, f_9, f_1, f_7\}$ |
| 9 | $\{f_1, f_3, f_5, f_8, f_2, f_4, f_6, f_7, f_9\}$ | $\{f_9, f_4, f_3, f_5, f_8, f_1, f_2, f_6, f_7\}$ | $\{f_3, f_4, f_5, f_8, f_2, f_9, f_1, f_7, f_6\}$ | $\{f_3, f_5, f_4, f_8, f_2, f_9, f_1, f_7, f_6\}$ |

TABLE III. MISCLASSIFIED INSTANCES WITH INDEX VALUES AND CORRESPONDING HEALTH METRICS

| Index | Age | BMI | Glucose | Insulin | HOMA | Leptin | Adiponectin | Resistin | MCP-1 | Actual Class | Predicted |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 50 | 38.5 | 106 | 6.7 | 1.7 | 46.6 | 4.6 | 11.7 | 887.1 | 1 | 2 |
| 100 | 74 | 28.6 | 88 | 3.0 | 0.65 | 31.1 | 7.6 | 18.3 | 572.4 | 2 | 1 |
| 31 | 53 | 36.8 | 101 | 10.1 | 2.5 | 27.1 | 20.0 | 10.2 | 695.7 | 1 | 2 |
| 55 | 34 | 24.2 | 92 | 21.7 | 4.9 | 16.7 | 21.8 | 12.0 | 481.9 | 2 | 1 |
| 114 | 72 | 25.6 | 82 | 2.8 | 0.57 | 24.96 | 33.75 | 3.27 | 392.4 | 2 | 1 |
| 30 | 66 | 36.2 | 101 | 15.5 | 3.87 | 74.7 | 7.5 | 22.3 | 864.97 | 1 | 2 |

*1) Mutual information:* The results show how adding metabolic and inflammatory markers affects the model's ability to make predictions. Initially, we have used only Age ($C_{mi1}$). Then the model has given moderate results with 65.7% accuracy and an F1-score of 62.5%. Interestingly, adding Glucose ($C_{mi2}$) and HOMA ($C_{mi3}$) did not improve performance. We have observed a significant gain with ($C_{mi4}$) when Resistin was added. It gives the accuracy to 71.4% and the F1-score to 72.2%. This means that Resistin helps improve the model's ability to differentiate between cases. The inclusion of BMI ($C_{mi5}$) maintained performance, while the addition of Insulin ($C_{mi6}$) has resulted in the highest accuracy (74.3%) and F1-score (74.3%). However, adding Leptin ($C_{mi7}$) and Adiponectin ($C_{mi8}$) slightly decreases the performance of the model. Finally, MCP-1 ($C_{mi9}$) recovered some of the lost performance. Then the accuracy is 71.4%.

*2) Chi-square score:* The results illustrate how the sequential addition of metabolic and inflammatory markers influences model performance using the Chi2 method. Initially, we used only MCP-1 ($C_{\chi 1}$), which gave an accuracy of 54.3% and an F1-score of 27.3%. This baseline performance was quite good. Then we have included Insulin ($C_{\chi 2}$). The inclusion of Insulin has decreased the performance with the accuracy of 37.1% and the F1-score to 21.4%. We have observed a significant improvement when we included Glucose ($C_{\chi 3}$). Then the accuracy jumped to 62.9% and the F1-score also reached 62.9%. This indicates that Glucose has a very important role for this classification. When we have added HOMA in ($C_{\chi 4}$), it gives a boost in the F1-score to 64.9% but accuracy remains quite similar. The addition of Resistin ($C_{\chi 5}$) further improved performance, increasing accuracy to 65.7% and the F1-score to 68.4%. The most substantial performance gain occurred when Age was included ($C_{\chi 6}$). It has increased the accuracy up to 80% and the F1-score to 78.8%. This shows that Age is a strong predictor in combination with the other markers. After adding BMI ($C_{\chi 7}$), the highest performance is maintained. The inclusion of Leptin ($C_{\chi 8}$) led to a slight decrease in

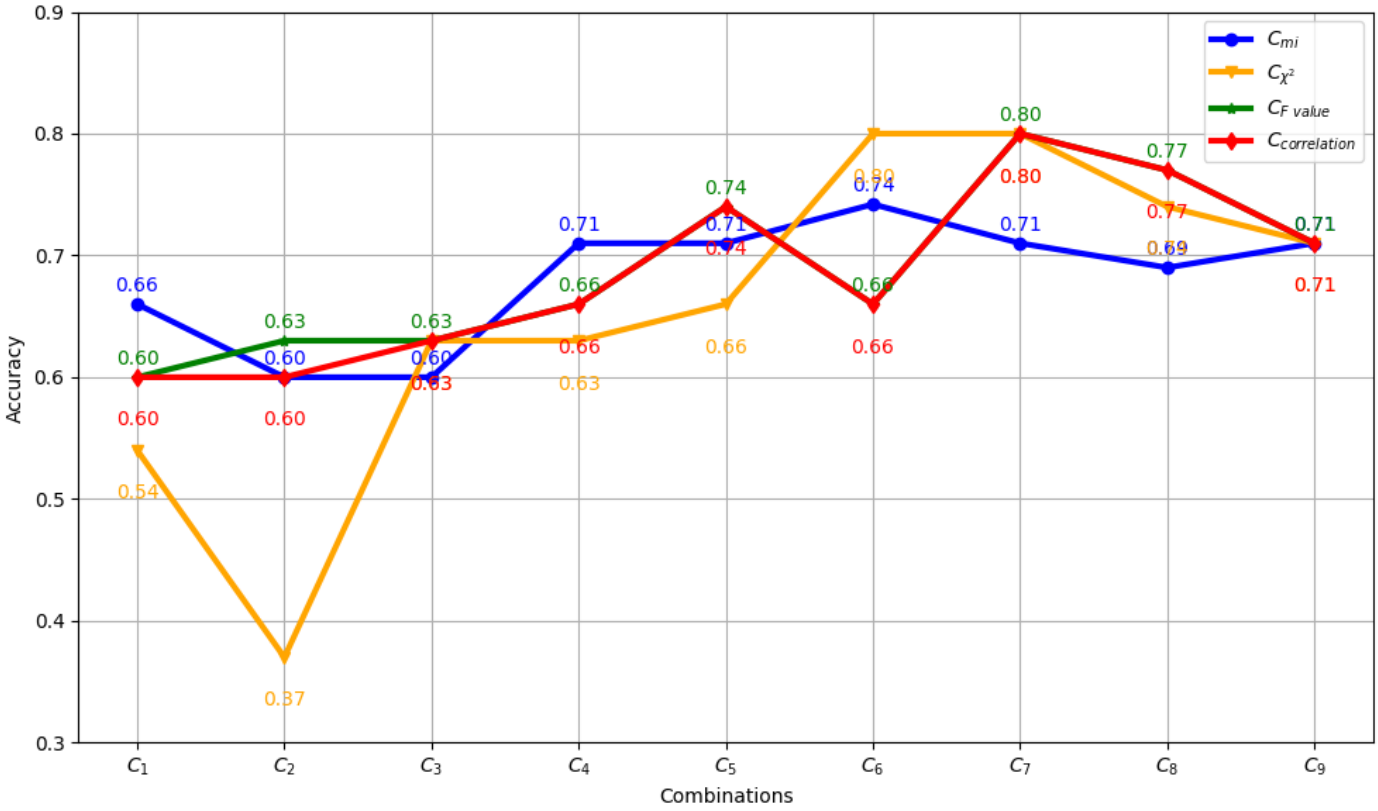


Fig. 5. Stepwise feature addition performance using mutual information score, chi-square value, ANOVA F-statistic, and correlation coefficient methods.

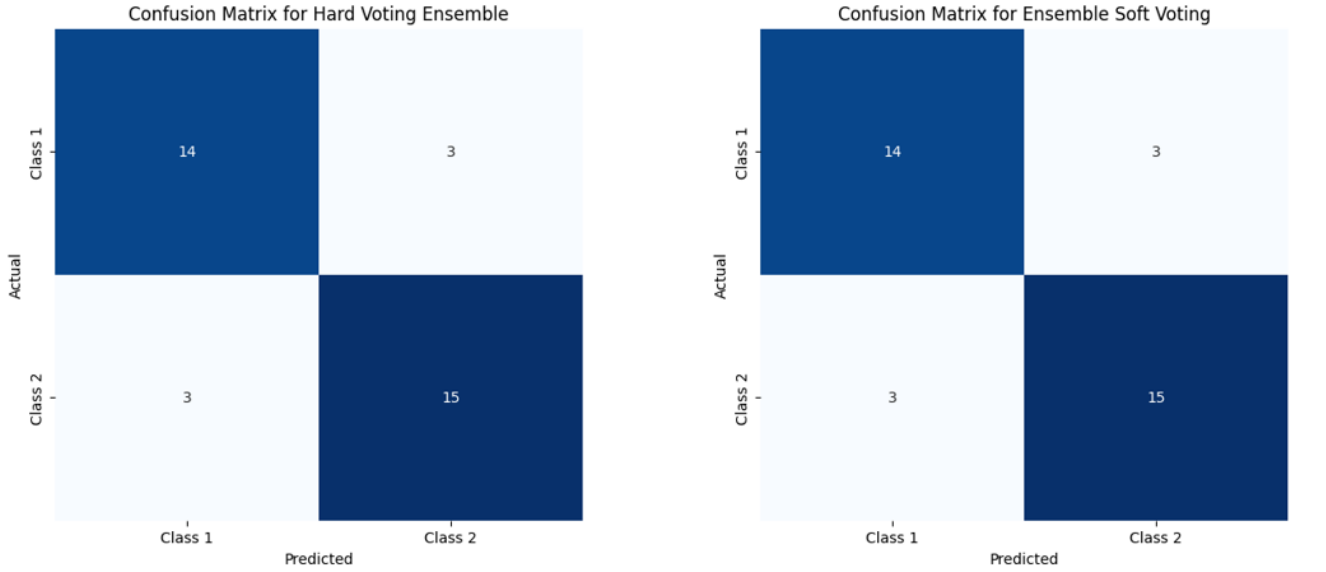


Fig. 6. Confusion matrix both for hard and soft voting.

performance, with accuracy dropping to 74.3% and the F1-score to 74.3%. Finally, adding Adiponectin ($C_{\chi 9}$) has resulted in a further decrease, with accuracy to 71.4% and the F1-score at 70.6%. This indicates that Adiponectin is not an important biomarker in this study.

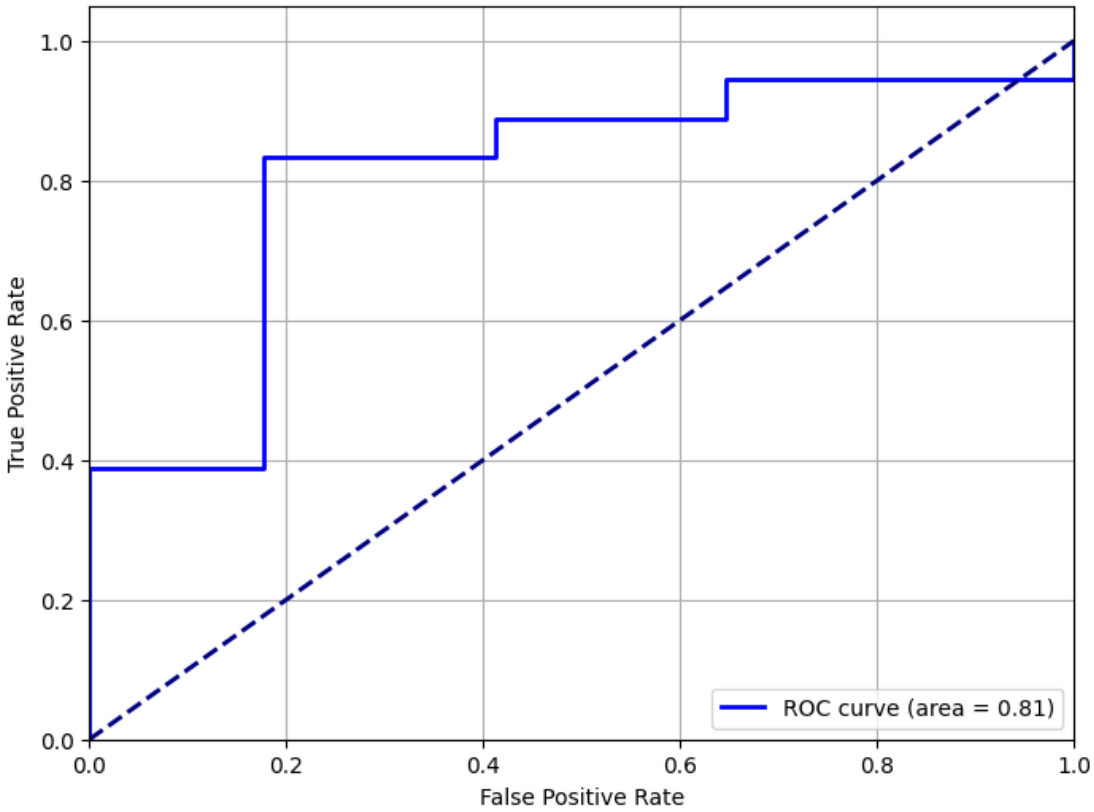


Fig. 7. Receiver Operating Characteristics (ROC) curve.

*3) ANOVA f-value:* The results show how progressively adding features impacts model performance using ANOVA-based selection. Initially, we have used only Glucose according to the ANOVA f-value which gives the accuracy of 60% and an F1-score of 58.8%. Then we have added Insulin ($C_{\text{f-value2}}$). It slightly improves the model performance, increasing both accuracy and F1-score to approximately 62.9%. After including HOMA ($C_{\text{f-value3}}$), it does not change the performance. However, when Resistin is added ($C_{\text{f-value4}}$), the model's performance improves to an accuracy of 65.7% and an F1-score of 68.4%. A significant boost is observed with the inclusion of BMI ($C_{\text{f-value5}}$). Then the accuracy rises to 74.3% and the F1-score to 76.9%. In this stage, the recall was 88%. When we have included MCP-1 ($C_{\text{f-value6}}$), the performance dropped. The addition of Age ($C_{\text{f-value7}}$) recovers the accuracy. Then the accuracy was found 80% and F1-score was 78.8%. When Adiponectin is added ($C_{\text{f-value8}}$), performance slightly decreases but remains robust (accuracy 77.1%, F1-score 75%). Finally, incorporating Leptin ($C_{\text{f-value9}}$) leads to a further small decline, stabilizing at 71.4% accuracy and an F1-score of 70.6%. This progression shows that some features give useful information and improve the model. But after a point, adding more features gives less benefit or can even hurt the model's accuracy.

Table II shows the ranking of features using the four selection methods. Glucose, HOMA, Insulin, Resistin, and Age appear as the most important features for breast cancer prediction.

### A. Ensemble of Classifiers Trained on Distinct Feature Subsets

In this study, an ensemble voting classifier has been constructed using four pipelines. Each pipeline was based on a distinct feature selection method: mutual information, chi-square, ANOVA F-value, and correlation coefficient. The selected feature sets for these methods are as follows:

$C_{mi6}$ = ['Age', 'Glucose', 'HOMA', 'Resistin', 'BMI', 'Insulin']

$C_{\chi 6}$ = ['MCP.1', 'Insulin', 'Glucose', 'HOMA', 'Resistin', 'Age']

$C_{\text{f-value7}}$ = ['Glucose', 'Insulin', 'HOMA', 'Resistin', 'BMI', 'MCP.1', 'Age']

$C_{\text{correlation7}}$ = ['Glucose', 'HOMA', 'Insulin', 'Resistin', 'BMI', 'MCP.1', 'Age']

Each pipeline used the AdaBoost classifier to enhance performance. The ensemble model combined predictions from the four AdaBoost-based pipelines using both hard voting and soft voting strategies. Both methods gave the same results, showing that the classifiers are well aligned. The ensemble achieved an accuracy of 82.86%, with F1-score, precision, and recall all measured at 82.35%. In Fig. 6, the confusion matrix shows that the model correctly predicted 14 instances of Class 1 and 15 instances of Class 2, with only three misclassifications for each class. This balance shows that the model performs well on both classes. The ROC curve (Fig. 7) depicts the true positive rate (TPR) versus false positive rate (FPR) across thresholds, with an AUC of 0.81 indicating fair separability.

### B. Misclassified Instances

Misclassification happens when the model predicts the wrong class for a data point. In this case, the Table III shows 6 misclassified instances (indices 26, 100, 31, 55, 114, 30). These are the cases where the actual class and predicted class do not match. The main reason for misclassifications is that many features (like age, BMI, glucose) overlap between classes. This makes it hard for the model to clearly separate class 1 and class 2. Also, the model may not capture complex relationships between features. There might be some noise or variability in biological measurements.

The results section summarizes all the numerical findings, showing that the ensemble model performed consistently and well-balanced.

## IV. Discussion

The experimental outcomes shows that combination of feature selection with AdaBoost-based ensemble voting significantly improves prediction accuracy. The top-rated features—Glucose, HOMA, Insulin, Resistin, and Age—are uniformly identified by all the methods. These biomarkers are of biological significance. These findings emphasize that the machine learning outcomes conform to current medical evidence, thus ensuring the validity of the suggested method.

The ensemble voting model achieved balanced accuracy for both classes, eliminating bias toward either non-cancer or cancer samples. The use of four statistical feature selection methods guaranteed that relevant and redundant variables were eliminated, increasing model interpretation and accuracy. However, some class overlaps, evidenced by feature distributions, indicate that there are biomarkers with very low discriminative power. This biological overlap explains the low number of misclassified instances indicated in the results.

Although the model is predictive, it was only based on a relatively small data set (116 cases). This restricts its use in larger populations.

It remains for future studies to advance this research by validating the model against larger, multi-center data sets and adding it to imaging or genomics to enhance diagnostic accuracy.

## V. Conclusion

In this study, we applied AdaBoost and ensemble voting classifiers combined with feature selection methods to classify breast cancer cases using the Breast Cancer Coimbra dataset. By using four statistical filter techniques — Mutual Information, Chi-Square, ANOVA F-test, and Correlation Coefficient — we identified the most relevant features for prediction. Key features such as glucose, HOMA, insulin, resistin, and age were common across all selection methods. This highlights their strong predictive value and importance in breast cancer classification.

Our ensemble model reached an impressive 82.9% accuracy, with well-balanced precision, recall, and F1-scores. This shows it has strong and reliable predictive power in classifying breast cancer cases. Despite this, a few misclassifications were observed, mainly due to overlapping feature distributions and possible biological variability. These results show that combining feature selection with ensemble methods improves classification performance.

From the biomedical perspective, this study illustrates how the features glucose, HOMA, insulin, and resistin are highly correlated with breast cancer. These properties illustrate how metabolic and inflammatory problems are connected to the cancer. Because the biomarkers can be measured by simple blood tests, this procedure is a cheaper and non-invasive way to detect breast cancer at an early stage, especially where high-level medical imaging is unavailable.

From the AI perspective, this work illustrates that employing a lot of feature selection algorithms under an ensemble setting makes the system more robust and interpretable despite having little data. Both ensemble voting and AdaBoost techniques allow the model to work effectively without confusing it regarding which features to prioritize.

In the future,Utilizing the model on larger datasets from different hospitals will also make it stronger and more reliable. Overall, this study bridges medical utility and artificial intelligence and shows that interpretable ensemble learning is a useful and low-cost tool for early breast cancer detection.

## Acknowledgments

This research has been supported by the Research Management Center, Multimedia University.

## References

[1] National Cancer Institute, "What is cancer?" 2022, accessed: 2025-06-12. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[2] World Health Organization, "Cancer," 2020, accessed: 2024-06-01. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[3] World Cancer Research Fund, "Breast cancer statistics," 2022, accessed: 2024-06-01. [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics/

[4] World Health Organization, "Breast cancer," 2022, accessed: 2024-06-01. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[5] W. H. Organization, "Breast cancer now most common form of cancer: Who takes action," *World Health Organization*, 2021, available at: https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-takes-action.

[6] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[7] National Cancer Institute, "Breast cancer screening," 2024, accessed: 2025-06-01. [Online]. Available: https://www.cancer.gov/types/breast/patient/breast-screening-pdq

[8] J. Smith and L. Brown, "Addressing the rural cancer care gap," *Journal of Rural Health*, vol. 37, no. 4, pp. 615–620, 2021.

[9] Y. Zhang, R. Proenca, M. Maffei, M. Barone, L. Leopold, and J. M. Friedman, "Positional cloning of the mouse obese gene and its human homologue," *Nature*, vol. 372, no. 6505, pp. 425–432, 1994.

[10] J. M. Friedman, "20 years of leptin: Leptin at 20: An overview," *Journal of Endocrinology*, vol. 223, no. 1, pp. T1–T8, 2014.

[11] C.-J. Lee, M. Fu, X. Liu, and Y.-C. Ko, "Resistin promotes angiogenesis in endothelial cells: involvement of the pi3k/akt pathway," *Biochemical and Biophysical Research Communications*, vol. 423, no. 1, pp. 64–69, 2012.

[12] C. J. Bailey and R. C. Turner, "Insulin and cancer risk: Epidemiological and biological perspectives," *Nature Reviews Endocrinology*, vol. 16, no. 7, pp. 393–406, 2020.

[13] D. F. Hayes, C. Isaacs, and V. Espina, "Insulin, insulin-like growth factors and breast cancer risk," *Breast Cancer Research*, vol. 18, no. 1, pp. 1–8, 2016.

[14] L. Mathews and J. Smith, "Role of homa in assessing insulin resistance and beta-cell function in breast cancer patients," *Journal of Clinical Endocrinology*, vol. 105, no. 4, pp. 1234–1241, 2020.

[15] H. Kim, S. Lee, and J. Park, "Monocyte chemoattractant protein-1 (mcp-1) and its role in breast cancer metastasis," *Cancer Immunology Research*, vol. 9, no. 7, pp. 789–797, 2021.

[16] Y. Hernández-Julio *et al.*, "Breast cancer prediction based on clustering and pivot table techniques," in *Proceedings of the International Conference (assumed)*, 2018, accuracy: 95.90%.

[17] B. K. Singh, "Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 393–409, 2019.

[18] K. Polat and U. Senturk, "A novel ml approach to prediction of breast cancer: Combining of mad normalization, kmc based feature weighting and adaboostm1 classifier," in *2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, 2018, pp. 1–5.

[19] S. B. Akben, "Determination of the blood, hormone and obesity value ranges that indicate the breast cancer, using data mining based expert system," *IRBM*, vol. 40, no. 6, pp. 355–360, 2019.

[20] M. R. Islam and Y.-T. Poly, "Breast cancer prediction using k-nn classifier on coimbra dataset," in *Assumed Proceedings*, 2020, accuracy: 86.00%.

[21] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[22] D. Dua and C. Graff, "Uci machine learning repository [breast cancer coimbra data set]," http://archive.ics.uci.edu/ml, 2017.

[23] R. A. Fisher, *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

[24] K. Pearson, "Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.