Exploring Feature Selection for Microarray Classification

Muhammad Zaky Hakim Akmal[1], Devi Fitrianah[2]

[1 2]*Computer Science Department BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia*

E-mail address: muhammad.akmal003@binus.ac.id, devi.fitrianah@binus.edu

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: This study delves into the critical role of feature selection in enhancing the accuracy of microarray data classification, particularly in the context of ovarian cancer detection. By harnessing the power of machine learning techniques and microarray technology, the research endeavors to identify subtle gene expression patterns that serve as indicators of ovarian cancer. By leveraging machine learning techniques and microarray technology, subtle gene expression patterns indicative of ovarian cancer can be identified. The research explores the utilization of Principal Component Analysis (PCA) for dimensionality reduction and compares the effectiveness of feature selection techniques such as Artificial Bee Colony (ABC) and Sequential Forward Floating Selection (SFFS). The dataset used in this study comprises of 15154 genes, 253 instances, and 2 classes related to ovarian cancer. Through a comprehensive analysis, the study aims to optimize the classification process and improve the early detection of ovarian cancer. Moreover, the study presents the classification accuracy results obtained by PCA, ABC, and SFFS. While PCA achieved an accuracy of 96% and SFFS yielded a classification accuracy of 98%, ABC demonstrated the highest classification accuracy of 100%. These findings underscore the effectiveness of ABC as the preferred choice for feature selection in improving the classification accuracy of ovarian cancer detection using microarray data.

Keywords: Feature Selection, Microarray data, Machine Learning, Cancer Detection

1. INTRODUCTION

Cancer is one of the number one causes of death in the world that may appear in various parts of the human body, such as the lungs, heart, pancreas, and many other organs and tissues. Cancer may start forming due to abnormal cell growth that will spread out to other organs and tissues in the body. You can get cancer by smoking, drinking alcohol, being older, certain types of infections, and you can even get cancer just by getting older. The reason is because of the mutations of the tissue, and some of these mutations may contribute to the birth of cancers [1].

Based on the Worldwide Cancer Research Fund International, cancer remains a concern to the world, there will be an estimate of 18.1 million cancer cases in the world in 2020 alone. Among these cases, the most common cancers that people may have are breast and lung cancer, while pancreatic and ovarian cancer are considered types of cancers that are considered hard to detect until they have reached an advanced stage. Ovarian cancer is considered one of the deadliest cancers challenging to detect in its preliminary stages, leading to delayed diagnoses and potentially poorer outcomes. The tricky progression of ovarian cancer underscores the critical need for improved diagnostic methods capable of identifying the disease at its earlier stage when treatment options are most effective. Traditional diagnostic approaches, including physical examinations, cancer screenings, blood tests, and laboratory analyses, may not always be sufficient in detecting ovarian cancer, given its tendency to manifest with vague or nonspecific symptoms until it has reached an advanced stage. Furthermore, ovarian cancer poses a significant challenge in early detection and treatment due to its elusive symptoms until it reaches advanced stages. Machine learning techniques coupled with microarray technology offer a promising approach to address this challenge.

By analyzing gene expression patterns from microarray data, machine learning algorithms can identify subtle signatures indicative of ovarian cancer. By leveraging machine learning, researchers can go through massive datasets to identify molecular signatures of ovarian cancer, even in its earliest stages. These approach helps doctors understand the disease better and treat it sooner. Also, combining machine learning with microarray technology means doctors can give treatments that fit each person's unique situation, making treatments work better and causing fewer problems.

There are numerous studies related to predicting ovarian cancer using machine learning approaches, such as XGBoost [2], Softmax Discriminant Algorithm (SDA) [3], and Gradient Boosting Decision Tree [4]. All three journals

utilize ovarian cancer microarray data labeled as ‘normal’ and ‘cancer’. Microarray technology is a powerful tool used by scientists to study gene activity by comparing hundreds or even thousands of gene profiles between different conditions, such as healthy tissue and cancerous tissue. This method allows researchers to simultaneously monitor, identify, and understand thousands or even millions of gene patterns in a single experiment. However, the abundance of genes analyzed in microarray data results in high-dimensional datasets, which can pose challenges for analysis due to computational instability and what is known as the "Curse of Dimensionality."

To address these challenges, dimensionality reduction technique is used to reduce the high-dimension data and to reduce computational instability. One commonly used technique is Principal Component Analysis (PCA), which aims to reduce the dimensionality of the data while preserving its essential features. In the journals [5], [6], [7], and [8], PCA (Principal Component Analysis) is utilized to tackle the Curse of Dimensionality in Microarray data. These journals obtained poor values generated by PCA compared to other feature reduction techniques. Therefore, this study will employ a feature selection technique that can choose important features based on the evaluation model to be included in the classification using Artificial Neural Network (ANN). Researchers explored alternative approaches to dimensionality reduction and feature selection in the context of microarray data analysis. One of the technique is ABC (Artificial Bee Colony) Feature Selection in the journals [9], [10], and [11], which selects prominent features based on a colony concept, mimicking the behavior of real life honey bee collecting food. The other feature selection technique is Sequential Forward Floating Selection (SFFS) as a comparison for ABC will be implemented in this research. SFFS has gained attention as an effective feature selection technique for handling high dimensional microarray data such as in the journals [12], [13], and [14]. In this research, the authors suggested in comparing the performance of PCA for dimensionality reduction and comparing the feature selection technique of ABC and SFFS using an Ovarian Cancer dataset sourced from [15]. This dataset consisted of 15154 genes, 253 instances and 2 classes, providing a robust foundation for evaluating the effectiveness of different approach in the context of cancer detection. By comparing these method, the study aims to identify the most effective strategy for optimizing the analysis of microarray data.

2. RELATED WORKS

Cancer is a complex and multifaceted disease characterized by the uncontrolled growth and spread of abnormal cells in the body. These abnormal cells, known as cancer cells, have the ability to invade and destroy surrounding tissues and organs. Cancer can arise in virtually any part of the body and can manifest in various forms, depending on the type of cells affected and the location of the tumor. While the most tumorous lesions are typically categorized as either "benign" or "malignant," the classification of ovarian tumors follows a more nuanced categorization, including "benign," "borderline," or "malignant" distinctions. Ovarian tumors encompass a spectrum of growths ranging from non-cancerous (benign) to potentially cancerous (malignant), with some falling in an intermediate category referred to as borderline tumors. Compared to benign ovarian tumors, malignant ovarian cancers are relatively rare, though they pose a significant health risk due to their potential to spread to other parts of the body. Borderline tumors, while less common than benign tumors, also present unique challenges in diagnosis and treatment due to their ambiguous nature, exhibiting features that lie between benign and malignant tumors [16] A Study from [17] found that age and ovary tumor site were significantly correlated with patient survival in ovarian cancer (OC). The study also identified clinical factors such as American Indian, African American, patient age, and cancer stage status as associated with significantly more risk of death within 5 years in OC. Patients with left site tumor in the ovary had a lower risk of death The study provides strong evidence that these genes are important prognostic indicators of patient survival and give clues to biological processes underlying OC progression and mortality. The study identified several genes, including TLR4, BSCL2, CDH1, ERBB2, SCGB2A1, and BRCA2, that were independently related to survival in ovarian cancer (OC) patients. These genes were found to be important prognostic indicators of patient survival and provided mechanistic and predictive information in addition to clinical traits. Age and ovary tumor site were significantly correlated with patient survival in OC. Additionally, clinical factors such as American Indian, African American, patient age, and cancer stage status were associated with a higher risk of death within 5 years in OC. Another study from [18] conducted a research where there were 607 cases of cancer recurrence and 416 cases of overall death in the median follow-up of 47 months (range of 4 to 177 months) for the training cohort. The majority of patients had FIGO stage III (56.6%) and grade 3 disease (56.8%) of high-grade serous type (61.6%). No residual disease after initial debulking surgery was observed in 469 patients (41.5%). For the validation cohort, there were 143 cases of cancer recurrence and 81 cases of overall death in the median follow-up of 63 months (range of 6 to 143 months).

DNA Microarray is a technology that is used to detect and compare thousands of gene profile samples at the same time. The principle is based on the hybridization of nucleic acid sequences, allowing researchers to simultaneously analyze the expression levels of thousands of genes or detect specific genomic sequences. The dimensionality of microarray data often poses challenges in the development of machine learning and even deep learning models. Many

journals have discussed various techniques for addressing data dimensionality in microarrays, employing methods such as feature selection and dimensionality reduction. In previous studies [6], a combination of the U-Net Neural Network and Unsupervised Principal Component Analysis (PCA) algorithms is used for the segmentation of cancer nests from hyperspectral images of breast cancer tissue microarray samples. The PCA technique in this journal aims to reduce computational complexity and enhance accuracy in the segmentation process. Another journal [19], explores an analytical platform for gastric cancer using Surface-Enhanced Raman Scattering (SERS) and PCA-two-layer nearest neighbor. The combined PCA model yielded an accuracy of 97.5%, sensitivity exceeding 90%, and specificity 96.7%. In the third journal [20], PCA techniques are discussed to improve the accuracy of gastric cancer prediction and identify patterns and differences in samples from patients with and without gastric cancer.

PCA became one of the dimensionality reduction techniques that is widely used. While other feature selection that is rarely used becomes the main alternative in facing microarray data such as Artificial Bee Colony (ABC) and Sequential Forward Floating Selection. The drawback of using a microarray is that there are a lot of features that can cause a curse of dimensionality. Journal from [21] explains that the ABC algorithm has enormous potential and can be implemented as an evolutionary structure that integrates the parameters of various traditional or modern heuristic algorithms. One of those potentials is explained in [22] where it uses the explore features of ABC algorithm and uses the attacking feature of another algorithm named Whale Optimization Algorithm. Another example is where [23] proposes an integrated standard error-based solution search into the original ABC algorithm. Based on the various studies, the synergy between ABC algorithm and other heuristic approaches emerges as a potent strategy for tackling the large dimensions of DNA microarray datasets and the curse of dimensionality. In [14], SFFS is discussed as a feature selection technique in the modeling process. By selecting relevant feature subsets from the available set, SFFS helps achieve the goal of constructing a miRNA biomarker panel that can serve as an indicator for breast cancer. The journal [24] explores various feature selection techniques, including Filters, Wrappers, and Embedded Approaches. SFFS falls under the Wrapper approach, and the selected features are only considered when accuracy exceeds 80%. The data used in this journal is sourced from UCI Machine Learning medical data. Another journal addressing Filters, Wrappers, and Embedded Approaches and utilizing SFFS as one of its techniques is [13]. In this journal, not only microarray data is used, but the approach is also applied to text analysis, intrusion detection systems, and stream data analysis. The researchers in this journal propose a novel approach in feature selection techniques for healthcare, government sectors, network attack predictions, and other domains.

3. METHODS

This section provides a comprehensive overview of the deep learning algorithm, feature selection techniques, and dimensionality reduction methods employed in the study. The primary aim is to identify the most effective combination among the chosen techniques to achieve optimal accuracy in cancer detection. Each feature selection method undergoes a standardized process, as illustrated in Figure 1, ensuring consistency and comparability across all approaches. All features selection undergoes identical preprocessing steps to prepare the data for analysis. This includes data cleaning, normalization, and transformation to ensure uniformity and accuracy in subsequent analyses. All the dataset is divided into a training set (80%) and a testing set (20%) using their respective methods. The training sets are then used to train an ANN classifier specific to each feature selection technique. These classifiers are optimized to recognize patterns and relationships within the data, enhancing their predictive capabilities. Meanwhile, the testing sets mirror the selected features from their corresponding training sets and are used to evaluate each method's performance. The accuracy results are used for comparison. By assessing accuracy across different feature selection methods, researchers can determine the superior feature selection method for cancer detection.

A. Pre-processing

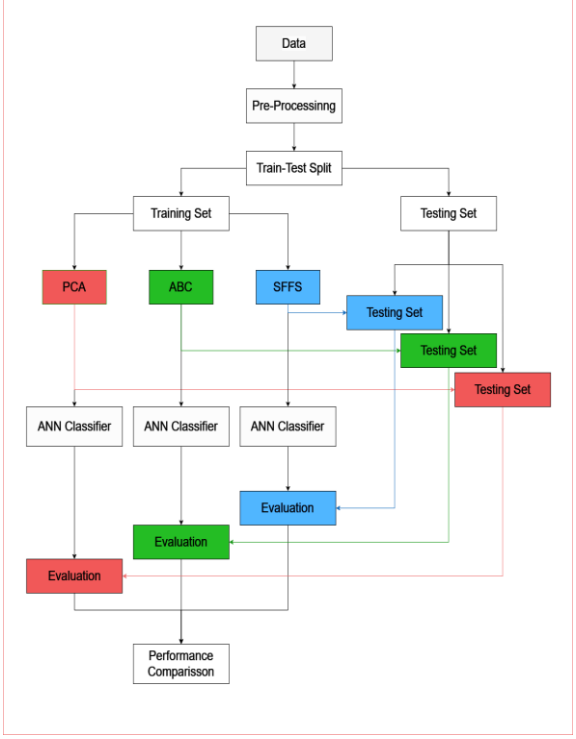

Figure 1. Research Framework

The dataset used in this research consists of 253 gene data points for each patient. After undergoing cleaning and data cleansing processes (checking for missing values and duplicate data), no problematic data was identified. In this preprocessing stage, the target feature is encoded by transforming the label 'Normal' into 0 and 'Cancer' into 1. Moreover, non-essential features such as patient ID will not be utilized in the modeling, so irrelevant features are dropped. Data normalization is performed on the training data with the aim of aiding the convergence of modeling algorithms more quickly and generating a better model [25]. The normalization step employs the Standard Scaler (1), which utilizes standard deviation for the data after the train-test split phase.

$$X_{new} = \frac{X_i - X_{mean}}{Standard\ Deviation} \quad (1)$$

B. Train-Test Split

Train-Test Split is a fundamental technique in machine learning that is used to evaluate the performance of the predictive models. It involves dividing a dataset into two subsets: one for training the model and the other for testing its performance with a ratio of 8:2. By allocating a majority of the data to the training set, the model sees many different examples, which helps it learn patterns, connection, and relationship in the data. Train-Test split is crucial for assessing a model's ability to generalize to new unseen data. It helps detect overfitting, where the model memorizes the training data to the extent that it performs poorly on new, unseen samples. Train-Test Split offers a robust mechanism for gauging the model's performance in real-world scenarios, mirroring its effectiveness in making predictions on data points. Furthermore, this technique furnishes an unbiased estimate of the model's performance, free from the biases that may arise from training and testing on the same dataset.

C. Principal Component Analysis (PCA)

PCA is a technique used in statistics and machine learning for dimensionality reduction and feature extraction. Its goal is to transform a high-dimensional dataset into a lower-dimensional space while retaining as fewest components as possible. PCA achieves this by trimming to keep the high value data and get rid of the rest, this will give a sense of complexity in the data set. Utilizing PCA for dimensionality reduction decreases the complexity of dimensions by allowing the microarray data to derive its features from eigenvectors and eigenvalues acquired during the process [26]. PCA is also flexible and can analyze datasets that contain missing values, categorical data, and unspecific measurements [7].

D. Artificial Bee Colony (ABC)

ABC is a population-based metaheuristic inspired from the metaphor of foraging behavior of honey bees in their quest for food. This algorithm encapsulates the essence of collaboration observed in the natural world, particularly among bees, to tackle the intricacies of solving complex problems across various domains. At the heart of the ABC algorithm is its iterative nature, where a series of phases occur to gradually optimize possible solutions and achieve the optimal result.

The process begins with an initialization phase. In this phase, the algorithm sets the stage by initializing a population of solutions, similar to starting a honey bee colony. Then, the employed phase begins, where bees actively explore the solution space and use local search mechanisms to find promising solutions. Following the employed phase, the onlooker phase takes center stage, reflecting the collective decision-making process observed as bystander bees evaluate and select solutions based on their quality and suitability. This phase embodies the essence of information sharing and collaboration, as onlooker bees exchange valuable insights to guide the collective pursuit of optimal solutions. After that, the ABC algorithm incorporate the scouting phase where the scout bees play a key role in identifying and replacing solutions that have reached stagnation or no longer hold promise. This phase adds dynamic elements to the algorithm, ensuring adaptability and resilience in the face of evolving problem situations. By seamlessly coordinating these phases of initialization, employment, onlooker, and scouting, ABC strikes the balance between exploration and exploitation, global and local search, and ultimately delivers unparalleled quality. Through the iterative process of exploration, exploitation, and information sharing, ABC converges towards optimal solutions by balancing local and global search [21].

E. Sequential Forward Floating Selection (SFFS)

SFFS is a wrapper feature selection method that will add one feature at a time to the selected set of features. At each iteration, the performance is evaluated using a chosen evaluation through cross-validation or another validation method. The feature with the highest performance will be added to the selected set [27]. During each iteration, SFFS identifies the features that yield the greatest performance improvement when added to the selected feature set. This feature is integrated into the set and increases its uniqueness. SFFS then dynamically evaluates the performance impact of feature removal. Excluding previously selected features improves performance, and SFFS selectively removes features if they indicate redundancy or noise in the feature set. This iterative process continues until no further improvement in performance is observed or a predefined stopping criterion

is met. By systematically exploring the feature space in this way, SFFS identifies the most informative and discriminatory subset of features for a given task, thereby maximizing prediction accuracy and other performance metrics. The purpose is that.

F. ANN-Classifier

ANN is one of the most used computational models of deep learning that is inspired by the way nerve cells work in the brain. Deep Learning automatically learns the data features to find complex patterns using multiple hidden layers of neural network to model and solve complex problems [28]. ANN consists of nodes that often converge into layers. The layers typically include an input layer, one or multiple hidden layers, and an output layer. Data will then enter the input layer and may pass through the hidden layer until it reaches the output layer [29]. Grid search, random search and KFold Cross-Validation are some of the most popular methods to be used to find the best number of units in an ANN hidden layer.

4. EXPERIMENT AND RESULT

A. Experiment using PCA

The PCA analysis begins by applying the preprocessing steps detailed in the methodology section. Once the data has been standardized through the standard scaler, the scaled dataset is utilized to determine the optimal number of components using PCA's explained variance ratio. By plotting the cumulative explained variance ratio against the number of components, the analysis identifies a threshold where the curve starts to level off. This inflection point indicates the optimal number of components to retain. Subsequently, this chosen number of components is pinpointed using a threshold, ensuring the most informative features are captured for further analysis.

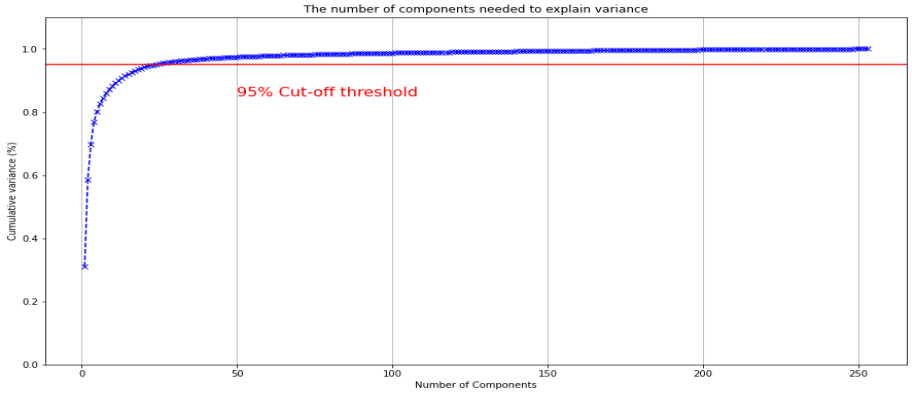

Figure 2. PCA Number of Components

The Threshold that is commonly used for PCA ranges from 95% to 99% to determine the level of variance to retain in the transformed data. In this study, a 95% threshold is employed, resulting in 24 components formed by PCA as the new features for modeling just as shown in Figure 2. This threshold selection process is used to allow for the reduction of the original dataset, consisting of approximately 15,130 original data points to 24 columns that can represent the original 15,154 data points. The implementation of PCA facilitates the dataset while retaining the essential information necessary for modeling. After that, KFold Cross-Validation is then used to find the best optimal number of units in the ANN classifier, which includes one hidden layer. This technique enables the most suitable architecture for the ANN model, enhancing its predictive performance. The optimal configuration obtain from KFold Cross-Validation is then utilize into the ANN classifier and ultimately yielded a test accuracy of 96.08% and a test loss of 0.1378. These performance metrics signify the effectiveness of PCA-based dimensionality reduction approach in facilitating accurate classification of the ovarian cancer dataset.

B. Experiment using ABC

TABLE I. ABC FEATURE SELECTED

Iteration	***50***			***100***		
nColony	10	20	30	10	20	30
Selected Feature	13300	7498	7637	7580	7658	7593
Accuracy (%)	96	100	98	96	100	100
Loss	0.3265	0.0003	0.0885	0.1050	0.0002	0.000001

The ABC experiment aims to perform feature selection on ovarian cancer data using varying parameters, specifically nColony values of 10, 20, and 30, with 50 and 100 iterations for each nColony setting as shown as table I. Approximately 50-87% of the features are selected from the original dataset containing 15,154 features. The experiment on ABC was conducted in two stages, stage 1 with 50 iterations and stage 2 with 100 iterations. When using 50 iterations, colonies of 10, 20, and 30 were formed, each resulting in different selected features. In the 10th colony, 13300 features were selected out of 15154 making it the highest features selected. In the 20th colony, 7498 features were selected being the lowest features out of all the iterations. While in the 30th colony only selected 7637 features. When 100 iterations were used in the ABC experiment on the data, the 10th colony yielded the fewest selected features compared to other colonies in its iteration, totaling 7580. While the 20th colony yielded the most selected features compared to other colonies in its iteration, totaling 7658, and the 30th colony yielded a total of 7593 selected features.

For nColony 10 parameters with both iterations, both gave test accuracy of around 96% and test loss 0.3265 and 0.1050, respectively. For nColony 20, both iterations achieved 100% accuracy, with test loss 0.0003 and 0.0002. Finally, for nColony 30, the 50 iteration run achieves 98% accuracy with a test loss of 0.0885, while at the 100 iteration run maintains 100% accuracy with a minimal test loss of 0.00001. Based on table I, it can be inferred that a higher number of iterations and nColony generally results in better accuracy and loss scores. However, there is an exception where for iteration 50, an nColony of 30 has relatively worse results compared to nColony of 20 since nColony 30 required more iterations to yield better results than nColony of 20. Finally, as for the number of selected features, it can be seen that an nColony of 10 is too little as it selected 13300 features in the 50th iterations compare to the 7580 features in the 100th iteration. This shows that an nColony that is too small may potentially result in too many irrelevant features being selected. Meanwhile in nColony 20 and 30 it can be observed that between the 50th and 100th iteration where the number of selected features have barely change, which means that it is already very close to the optimal number of features.

C. Experiment using SFFS

Similar to the ABC experiment, the SFFS approach utilizes ovarian cancer data for feature selection. However, unlike ABC, SFFS does not employ nColony but relies on a classifier alone as its estimator. In this study, Logistic Regression is utilized instead of an ANN, as the keras layer model is not compatible with SFFS. Additionally, to ensure compatibility with SFFS, the y_train data is flattened using the numpy ravel function, this is done so that the data for each element of the data corresponds to a single feature, making it easier to evaluate and select the feature. As a result of this experiment, SFFS resulted in an accuracy of 98.04%, with SFFS successfully selecting a total of 7,577 features. The test loss is recorded at 0.0473, indicating the effectiveness of the selected features in accurately classifying ovarian cancer data.

TABLE II. COMPARISON RESULT

Method	**Accuracy (%)**	**Loss**
PCA	96	0,1378
ABC(50,10)	96	0,3265
ABC(50,20)	100	0,0003
ABC(50,30)	98	0,0885
ABC(100,10)	96	0,105
ABC(100,20)	100	0,0002
ABC(100,30)	100	0,00001
SFFS	98	0.0473

Table II. illustrates the remarkable performance metrics of various feature selection techniques, with Artificial Bee Colony (ABC) method achieving the highest accuracy among all feature selection techniques that was examined. Notably, ABC attained an outstanding accuracy of 100%, surpassing both PCA and SFFS, which achieved accuracies of 96% and 98% respectively. This

remarkable result underscores the effectiveness of ABC in discerning crucial gene expression patterns indicative of ovarian cancer. Further analysis reveals that the exceptional accuracy of ABC can be attributed to specific parameter configurations. In particular, ABC iterations at 50 and 100, with colony sizes of 10, 20, and 30, were explored. Intriguingly, the configuration that yielded the 100% accuracy comprised ABC iterations at 50 and 100, with colony sizes of 20 and 30, respectively. These findings highlight the critical role of parameter optimization in achieving optimal performance with ABC, and highlighting the importance of fine-tuning parameters to maximize accuracy. The exceptional accuracy achieved by ABC not only underscores its potential as a robust feature selection technique but also signifies its utility in enhancing the classification process in microarray-based cancer detection. Such insights gleaned from this study contribute significantly to the ongoing efforts aimed at advancing early diagnosis and treatment strategies for ovarian cancer patients, ultimately leading to improved clinical outcomes and patient care.

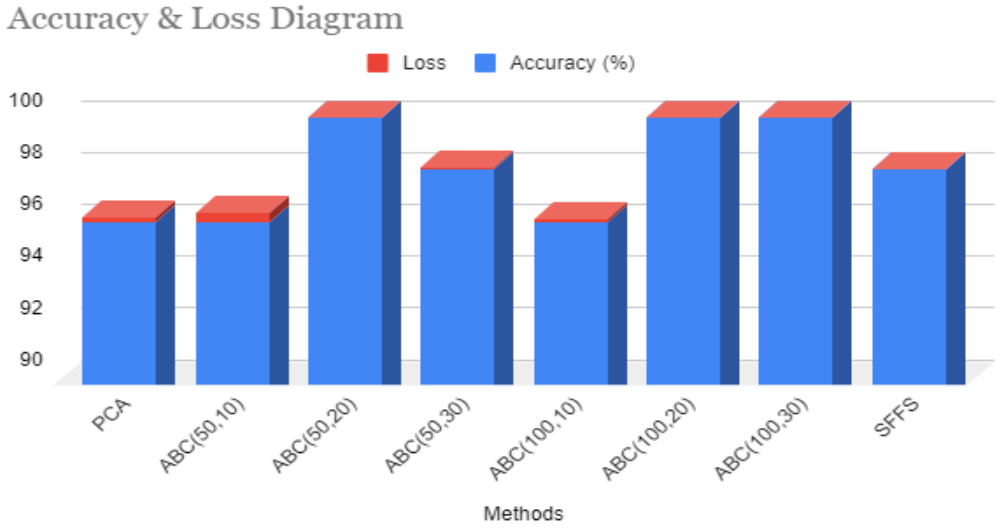

Figure 3. Accuracy & Loss Diagram

The comparative analysis of PCA, ABC, and SFFS reveals distinct approaches to feature selection and modeling in the context of ovarian cancer detection. Based on figure 3, PCA demonstrates its effectiveness by reducing the dataset's dimensionality to 24 components while maintaining a high accuracy of 96.08% through ANN modeling. Conversely, ABC, with its flexible parameter tuning and feature selection capabilities, achieves remarkable results, notably attaining a perfect 100% accuracy under optimal configurations. Meanwhile SFFS, although utilizing Logistic Regression due to compatibility constraints, efficiently selects 7,577 features with a high accuracy of 98.04%. However, it's important to note that SFFS had the longest running computational time among the three methods, that required more than a day to finish its computation, whereas PCA and ABC both took less than 8 hours combined. Despite this, each method showcases unique strengths: PCA offers simplicity and efficient dimensionality reduction, ABC excels in fine-tuning parameter configurations for optimal feature selection, and SFFS efficiently selects features with high accuracy, albeit with longer computational time. The selection among these approaches depends on several factors, including the characteristics of the dataset, available computational resources, and the specific objectives of the analysis. Researchers must carefully weigh these considerations to choose the most suitable method that aligns with their research goals and constraints. Moreover, further exploration and experimentation may be warranted to fully understand the nuances and trade-offs associated with each technique, ensuring robust and reliable results in the context of ovarian cancer detection and beyond.

5. CONCLUSION

In conclusion, the research underscores the critical role of feature selection in not only enhancing the accuracy but also optimizing the efficiency of microarray data classification for cancer detection, particularly in the challenging context of ovarian cancer detection. By employing advanced techniques and comparing them such as PCA for dimensionality reduction and feature selection methods like ABC and SFFS, the study demonstrates the potential for optimizing the classification process in the microarray dataset. From observing the accuracy, loss, and runtime values during a multiple number of experiments, it becomes evident that ABC provides more optimal results compared to PCA and SFFS. ABC, with its approach

inspired by the behavior of real bees in search of food sources, achieves a remarkable accuracy of 100% when using nColony size of 20 and demonstrate a minimum loss of 0.0003. Moreover, the runtime for implementing ABC require a manageable runtime, ranging around 1 to 3 hours for each of its experiment. On the other hand, PCA, while serving as a widely-used method for dimensionality reduction, yields relatively lower accuracy results compared |to ABC, emphasizing the need for more sophisticated feature selection approaches in microarray data analysis. Similarly, SFFS exhibits a significantly longer runtime, rendering it inefficient for microarray data usage in its current computational environment. However, it is noted that SFFS has potential to generate better outcomes when employed on a more powerful computing device, indicating the importance of considering hardware capabilities when selecting feature selection methods for complex datasets. Given its iterative nature and computational demands of SFFS, it benefits from enhanced processing power and memory resources, potentially unlocking its full capabilities in uncovering subtle gene expression patterns associated with ovarian cancer. These findings underscore the significance of harnessing machine learning algorithms and microarray technology to uncover subtle gene expression patterns associated with ovarian cancer. Such endeavors hold immense potential for advancing early detection and treatment strategies in cancer research, ultimately leading to improved patient outcomes and contributing to the broader effort of combating complex diseases. By continually refining and optimizing how computers analyze data and understanding how our bodies work, researchers can pave the way for transformative breakthroughs in the fight against cancer and other complex devastating illnesses, bringing hope to millions around the world.

6. FUTURE WORK

While this study has shed light on the critical role of feature selection in enhancing the accuracy and efficiency of microarray data classification for ovarian cancer detection, there remain several avenues for further exploration and refinement. Future work in this field could explore alternative feature selection techniques beyond those compared in this study, while this study compared PCA, ABC, and SFFS, there are numerous other feature selection methods that are available such as GA (Genetic Algorithm), Random Forest, or recursive feature elimination, to identify the most suitable approach for ovarian cancer detection.

Moreover, future work could explore beyond traditional metrics, such as accuracy and loss, to incorporate a much wider array of evaluation measures that provide a more comprehensive assessment of model performance and generalizability. By incorporating a diverse range of evaluation metrics, researchers can gain deeper insights into the strengths and limitations of different feature selection techniques and machine learning algorithms.

While the findings of this study are promising, it is essential to validate the result on independent dataset to ensure the robustness and generalizability of the proposed methods. Future studies could involve collaborating with multiple research institutions to access a more diverse dataset, facilitating external validation and replication of the findings. In conclusion, there are numerous opportunities for future research to enhance diagnostic accuracy.

Additionally, future research could extend the application of the ABC algorithm beyond microarray data analysis and ovarian cancer detection. Given its demonstrated effectiveness in handling high-dimensional datasets, ABC hold promise for improving diagnostic accuracy in the identification of other deadly disease characterized by complex genetic signatures and large feature. Other type of diseases such as pancreatic cancer and leukemia often present similar challenges in terms of data dimensionality and feature complexity. Exploring the applicability of ABC in these contexts could involve adapting the algorithm to suit the unique characteristics of each disease's molecular profile.

Furthermore, future studies could explore the integration of the use of other machine learning classification techniques other than ANN to enhance its performance and scalability in analyzing diverse disease datasets. With an example of using hybrid algorithms combining ABC with deep learning architectures, ensemble methods, or network-based approaches could offer synergistic advantages in capturing complex disease dynamics and improving predictive accuracy.

7. REFERENCE

[1] E. Laconi, F. Marongiu, and J. DeGregori, "Cancer as a disease of old age: changing mutational and microenvironmental landscapes," *Br J Cancer*, vol. 122, no. 7, pp. 943–952, 2020, doi: 10.1038/s41416-019-0721-1.

[2] N. B. Shannon *et al.*, "A machine learning approach to identify predictive molecular markers for cisplatin chemosensitivity following surgical resection in ovarian cancer," *Sci Rep*, vol. 11, no. 1, pp. 1–10, 2021, doi: 10.1038/s41598-021-96072-6.

[3] M. Kalaiyarasi and H. Rajaguru, "Performance Analysis of Ovarian Cancer Detection and Classification for Microarray Gene Data," *Biomed Res Int*, vol. 2022, 2022, doi: 10.1155/2022/6750457.

[4] K. Chen *et al.*, "Integration and interplay of machine learning and bioinformatics approach to identify genetic interaction related to ovarian cancer chemoresistance," *Brief Bioinform*, vol. 22, no. 6, pp. 1–11, 2021, doi: 10.1093/bib/bbab100.

[5] E. Lotfi and A. Keshavarz, "Gene expression microarray classification using PCA-BEL," *Comput Biol Med*, vol. 54, pp. 180–187, 2014, doi: 10.1016/j.compbiomed.2014.09.008.

[6] J. Wang *et al.*, "PCA-U-Net based breast cancer nest segmentation from microarray hyperspectral images," *Fundamental Research*, vol. 1, no. 5, pp. 631–640, 2021, doi: 10.1016/j.fmre.2021.06.013.

[7] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021, doi: 10.30880/jscdm.2021.02.01.003.

[8] E. Nazari, M. Aghemiri, A. Avan, A. Mehrabian, and H. Tabesh, "Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data," *Gene Rep*, vol. 25, 2021, doi: 10.1016/j.genrep.2021.101419.

[9] R. M. Aziz, "Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data," *Med Biol Eng Comput*, vol. 60, no. 6, pp. 1627–1646, 2022, doi: 10.1007/s11517-022-02555-7.

[10] E. H. Houssein, D. S. Abdelminaam, H. N. Hassan, M. M. Al-Sayed, and E. Nabil, "A Hybrid Barnacles Mating Optimizer Algorithm with Support Vector Machines for Gene Selection of Microarray Cancer Classification," *IEEE Access*, vol. 9, pp. 64895–64905, 2021, doi: 10.1109/ACCESS.2021.3075942.

[11] A. Jahwar and N. Ahmed, "Swarm Intelligence Algorithms in Gene Selection Profile Based on Classification of Microarray Data: A Review," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 01–09, 2021, doi: 10.38094/jastt20161.

[12] H. Younis, M. W. Anwar, M. U. G. Khan, A. Sikandar, and U. I. Bajwa, "A New Sequential Forward Feature Selection (SFFS) Algorithm for Mining Best Topological and Biological Features to Predict Protein Complexes from Protein–Protein Interaction Networks (PPINs)," *Interdiscip Sci*, vol. 13, no. 3, pp. 371–388, 2021, doi: 10.1007/s12539-021-00433-8.

[13] G. Manikandan and S. Abirami, "Feature Selection Is Important: State-of-the-Art Methods and Application Domains of Feature Selection on High-Dimensional Data," *EAI/Springer Innovations in Communication and Computing*, pp. 177–196, 2021, doi: 10.1007/978-3-030-35280-6_9.

[14] R. Zou *et al.*, "Development and validation of a circulating microRNA panel for the early detection of breast cancer," *Br J Cancer*, no. April 2021, 2022, doi: 10.1038/s41416-021-01593-6.

[15] Z. Zhu, Y. S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognit*, vol. 40, no. 11, pp. 3236–3248, 2007, doi: 10.1016/j.patcog.2007.02.007.

[16] M. Akazawa and K. Hashimoto, "Artificial intelligence in ovarian cancer diagnosis," *Anticancer Res*, vol. 40, no. 8, pp. 4795–4800, Aug. 2020, doi: 10.21873/anticanres.14482.

[17] E. S. Paik *et al.*, "Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods," *J Gynecol Oncol*, vol. 30, no. 4, Jul. 2019, doi: 10.3802/jgo.2019.30.e65.

[18] M. A. Hossain, S. M. Saiful Islam, J. M. W. Quinn, F. Huq, and M. A. Moni, "Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality," *J Biomed Inform*, vol. 100, Dec. 2019, doi: 10.1016/j.jbi.2019.103313.

[19] D. Cao *et al.*, "PCA-TLNN-based SERS analysis platform for label-free detection and identification of cisplatin-treated gastric cancer," *Sens Actuators B Chem*, vol. 375, 2023, doi: 10.1016/j.snb.2022.132903.

[20] L. Guo *et al.*, "Identification and analysis of serum samples by surface-enhanced Raman spectroscopy combined with characteristic ratio method and PCA for gastric cancer detection," *J Innov Opt Health Sci*, vol. 12, no. 2, pp. 1–11, 2019, doi: 10.1142/S1793545819500032.

[21] ABHISHEK SHARMA, ABHINAV SHARMA, SACHI CHOUDHARY, RUPENDRA KUMAR PACHAURI, AAYUSH SHRIVASTAVA, and DEEPAK KUMAR, "a Review on Artificial Bee Colony and Its Engineering Applications," *Journal of Critical Reviews* , vol. 7, no. 11, pp. 4097–4107, 2020, doi: http://www.jcreview.com/fulltext/197-1596854993.pdf?1597550239.

[22] P. Stephan, T. Stephan, R. Kannan, and A. Abraham, "A hybrid artificial bee colony with whale optimization algorithm for improved breast cancer diagnosis," *Neural Comput Appl*, vol. 33, no. 20, pp. 13667–13691, Oct. 2021, doi: 10.1007/s00521-021-05997-6.

[23] K. Hanbay, "A new standard error based artificial bee colony algorithm and its applications in feature selection," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4554–4567, 2022, doi: 10.1016/j.jksuci.2021.04.010.

[24] S. Bashir, I. U. Khattak, A. Khan, F. H. Khan, A. Gani, and M. Shiraz, "A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/8190814.

[25] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl Soft Comput*, vol. 97, 2020, doi: 10.1016/j.asoc.2019.105524.

[26] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification," *Journal of Computer Science*, vol. 14, no. 11, pp. 1521–1530, 2018, doi: 10.3844/jcssp.2018.1521.1530.

[27] B. E and J. Akpajaro, "Genetic Algorithm With Bagging for Dna Classification," *International Journal of Advances in Signal and Image Sciences*, vol. 7, no. 2, pp. 31–39, 2021, doi: 10.29284/ijasis.7.2.2021.31-39.

[28] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021, doi: 10.1007/s12525-021-00475-2.

[29] A. Shah *et al.*, "A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN)," *Clinical eHealth*, vol. 6, pp. 76–84, 2023, doi: 10.1016/j.ceh.2023.08.002.

Muhammad Zaky Hakim Akmal A student of Master Computer Science in Bina Nusantara University. His research interest is in machine learning and deep learning. His email can be contact at muhammad.akmal003@binus.ac.id

Devi Fitrianah received the Bachelor's degree in Computer Science from Bina Nusantara University, Jakarta, Indonesia, in 2000, and the Master's degree in Information Technology and Ph.D. degree in Computer Science from the Universitas Indonesia, Depok, Indonesia, in 2008 and 2015, respectively. In 2014, she had a sandwich program at the Laboratory for Pattern Recognition and Image Processing and GIS (PRIPGIS Lab) Department of Computer Science, Michigan State University, East Lansing, Michigan, USA.