IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 July 2023, accepted 29 July 2023, date of publication 11 August 2023, date of current version 18 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3304330

RESEARCH ARTICLE

A Comparative Analysis of Hybridized Genetic Algorithm in Predictive Models of Breast Cancer Tumors

JOYCE A. AYOOLA AND TOKUNBO OGUNFUNMI, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, Santa Clara University, Santa Clara, CA 95053, USA

Corresponding author: Joyce A. Ayoola (JAyoola@scu.edu)

This work was supported in part by the David Packard Endowed Faculty Fellowship, Santa Clara University.

ABSTRACT Advancement in computer-aided tools towards accurate breast cancer early prediction models has proven to be advantageous, which in turn helps to reduce the mortality rate associated with this cancer. From the literature, random forest predictor has been observed to have high accuracy in comparison to other machine learning regressors, also genetic algorithm has been observed to be a good feature selection method in data pre-processing. In a bid to improve the accuracy of breast cancer predictive models, several studies have developed hybridized genetic algorithm models for feature selection, however, the order of hybridization may not have been taken into consideration, as this can have an impact on the hybridized model's performance. Therefore, this paper proposes several high-performing predictive models using hybridized genetic algorithm, based on other learning models, while taking into consideration the placement order of the feature selection algorithms in the hybridized models. The Wisconsin Breast Cancer dataset was used as the test bench, while filter, wrapper and embedded feature selection algorithms were used in the proposed hybridized models. The performances of proposed hybridized models were compared with those of the individual learning models, considered in this work. These models include Fisher_Score, Mutual Information Gain, Correlation Chi-square test, Coefficient, Variance, Genetic Algorithm, Lasso and Linear Regressors with L1 regularization, Ridge Regressor with L2 regularization, Tree-based methods. From the performance evaluation results, the proposed hybridized Genetic Algorithm with Fisher_Score (GA + Fisher_Score) model showed promising results, as it had an accuracy score of 99.12%, thereby out-performing other proposed hybridized genetic algorithm models considered.

INDEX TERMS Random forest, genetic algorithm, breast cancer, prediction, feature selection, hybridization.

I. INTRODUCTION

Over the last few decades, Cancer has been known, universally, to be a deadly disease. In a 2022 USA cancer report by the American Cancer Society, heart disease was ranked as the most common disease closely followed by cancerous diseases. From the report, the total number of expected new cancer cases recorded was almost 2 million with over 609 thousand expected deaths [1]. The good news is that within the last 30 years, scientists have done tremendous work to help reduce the mortality rate associated with cancer disease. According to the American Cancer Society report [1], a significant decline in cancer related deaths was observed in the last 3 decades, to around 3.5 million fewer cases than expected.

One of the most common cancers, in the year 2022, is breast cancer which accounts for 12.5% of all new annual cancer cases, globally. From the American Cancer Society 2022 report, the most often diagnosed cancer cases recorded, among American women, was Breast Cancer. At least, 1 in every 8 American women will experience an invasive breast cancer during their life-time. The estimated number of new cases associated with breast cancer in 2022 was over 300 thousand women, with almost 289 thousand

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães.

invasive cases and a little over 51 thousand non-invasive cases [2].

The mortality rate associated with Breast Cancer has been observed to slowly decline over the years, to about 43% through 2020, with the help of treatment advancement and earlier detection. Nonetheless, more work still needs to be done to reduce the mortality rate thereby achieving the objective of enhancing health and well-being as set forth by the United Nations Sustainable Development Goals (SDG), as no definite cure or preventive measure has been discovered yet for breast cancer [3]. Detecting any abnormalities at an early stage before they develop into cancerous growth is vital, and mammogram screening and self-breast examinations play a crucial role in achieving this goal. If a lump is discovered during mammogram screening, it is classified as either a benign or malignant tumor. While benign tumors can be managed effectively and do not pose a significant threat, malignant tumors are highly aggressive and can invade nearby tissues rapidly.

Researchers and medical professionals have made tremendous advancement in the development of cancer-related predictive tools, which include the use of imaging data such as MRI (magnetic resonance imaging) and CT (computed tomography) scans to predict cancer by training computer vision algorithms to identify the presence of cancerous cells in medical images [4]. Also, researchers have combined several data sources such as patient demographic information, medical history, lab test results, and imaging data to develop more accurate breast cancer predictive models [5].

In recent times, researchers have incorporated machine learning to help classify and predict cancer, as computational techniques of machine learning algorithms, such as decision trees, random forests, support vector machines, and artificial neural networks, have been known to be effective [6]. Several works around machine learning have been conducted, gaining relevance to so many investigators, despite these advances, there are still challenges with cancer prediction such as limited data available for training predictive models which can cause overfitting or underfitting of models, leading to decreased accuracy. Predictive models can be biased if the training data reflects the experiences of only certain populations, such as those from specific geographic regions or ethnic groups. Other shortcomings include noisy information, redundancy, curse of dimensionality.

The motivation behind the proposed hybridization of genetic algorithm based on other learning models, stems from the need to improve model performance, interpretability, and generalization ability by selecting relevant and non-redundant features. The main challenges faced include handling high-dimensional datasets, preventing overfitting, exploring the large search space, striking a balance between feature relevance and redundancy, and ensuring scalability for large-scale datasets.

The remainder of the paper is organized as follows: Section II contains a review of related works; Section III presents the proposed methodology for feature selection; Section IV describes the performance evaluation of proposed methodology; in Section V the results are presented; in Section VI discussions were presented; finally, in Section VII conclusions are drawn while future work is suggested.

II. LITERATURE REVIEW

In a recent study conducted by the authors [7], Random Forest was observed to be a highly effective algorithm for breast cancer prediction and its performance was improved further when a feature selection technique, Genetic Algorithm, was included in its pre-processing phase. Feature selection is said to have relevance in machine learning as it is helpful in obtaining pertinent and germane data from a large multi-dimensional dataset, which may also help to reduce computational cost and improve the classification performance by taking out irrelevancies and noise from the data [8], [9], [10], [11]. In comparison to other dimensionality reduction techniques which make alterations to the initial data set, feature selection methods do not make any alterations [12]. Although much research has been done around feature selection, improvements are still needed, such as in computational complexity and noise reduction.

From the literature, we observed that hybridization of feature selection methods in the preprocessing phase of a classifier's model helps in improving the accuracy and stability of the classifier's predictive performance. Feature selection hybridization helps to overcome the limitations of individual methods, as different feature selection methods have their own strengths and limitations, hence combining multiple methods would lead to a more accurate and robust feature set [13]. Hybridization also helps to reduce overfitting and improve the generalization ability of predictive models by considering a wider range of feature selection criteria [14]. By combining multiple feature selection methods, hybridization can produce a feature set that is more interpretable, making it easier to understand how the features are related to the outcome. This can be useful in domains such as medical diagnosis, where understanding the underlying relationships between features and disease is important. High-dimensional data, such as data with a large number of features, can be difficult to handle using traditional feature selection methods. Hybridization can be used to reduce the dimensionality of the data and improve the performance of predictive models.

Kawamura and Chakraborty [15] combined two filter methods (Correlation and Minimum Redundancy-and-Maximum Relevance) with two wrapper methods (Binary Genetic Algorithm and Particle Swarm Optimization) for feature selection. They utilized Support Vector Machines (SVM) as the classifier on datasets from the UCI machine learning repository and two additional datasets. The experiments demonstrated that the hybridized model was effective in selecting optimal features, improving SVM accuracy, and reducing computational costs. Jain and Singh [16] proposed a hybridized framework that combined relief feature ranking with Principal Component Analysis (PCA) as a feature selection method, while k-nearest neighbor was used as

the classifier for breast cancer and diabetes diagnosis. The performance of their hybridized framework was compared with Mutual Information and four other feature selectors, with results showing that the proposed hybridized framework achieved over 80% classification accuracy and outperformed the other frameworks. Sahmadi and Boughaci in [17] proposed a hybridized genetic algorithm and simulated annealing meta-heuristic as a feature selection method, after which SVM classifier model was used on 11 datasets, including a breast cancer dataset. Before the hybridized method was used, SVM had an accuracy of 97.12% which was improved to 97.46% when the hybridized method was included.

We observe from the literature that further improvement of hybridized feature selection methods can increase accuracy and efficiency of predictive models even when applied to large datasets. Also, the robustness of predictive models increases while dependency on a single method reduces, improving their performance under different conditions. However, the order of hybridization may not have been taken into consideration, as this can have an impact on the hybridized model performance.

To this end, this paper hybridizes genetic algorithm with other feature selection methods in the data pre-processing phase of Random Forest classifier, with the aim to further advance the classifier's performance. The Feature selection categories used were filter, embedded and wrapper methods. For the wrapper method, only the genetic algorithm used in our previous study [7] was considered as this category of feature selection algorithms are computationally expensive, hence hybridizing two wrapper-based algorithms will be even more computationally expensive. Once the prediction stage of the hybridized models was completed, the models' predictive results were assessed and compared for performance evaluation.

III. PROPOSED METHODOLOGY

In this paper, three classes of feature selection techniques were examined, namely Filter methods (Fisher_Score, Mutual Information Gain, Correlation Chi-square test, Coefficient, Variance), Wrapper methods (Genetic Algorithm), and Embedded methods (Lasso and Linear Regressors with L1 regularization, Ridge Regressor with L2 regularization, Tree-based methods).

Filter methods assess feature relevance based on their correlation with the dependent variable. They are computationally faster than other methods as they do not require training models, and use statistical techniques to evaluate feature subsets. However, they may not always find the best subset of features. On the other hand, wrapper methods measure the usefulness of feature subsets by training models on them. They utilize cross-validation and can always provide the best feature subset, but they are computationally expensive due to repeated learning steps and cross-validation. Additionally, wrapper methods may increase the risk of overfitting the model as compared to using feature subsets from filter methods. Embedded methods and wrapper methods serve a similar purpose of enhancing the objective function or performance of a learning algorithm or model. However, unlike wrapper methods, embedded methods employ an inherent model building metric during the learning process.

Due to the computational expense of wrapper methods, we only considered the genetic algorithm from that category in the proposed hybridized models, which gave us an improved model performance in our previous study [7], Five (5) filters, and four (4) embedded methods were considered in the feature selection process. In this paper, the Wisconsin Hospitals Madison Breast Cancer Database [12] was used as the diagnostic dataset for breast cancer. The dataset comprises of 569 samples and 32 features, with complete data and a target variable of either Benign or Malignant. Out of the samples, 357 correspond to benign and 212 to malignant tumor outcomes. 80% of the dataset was used for training, while 20% was used for testing. To predict the occurrence of benign and malignant tumors in the dataset, the Random Forest classifier was employed. We carried out several combinations of hybridized models which were categorized into three cases. These three cases are three methodologies considered. For each methodology, several hybridizations, with genetic algorithm, were produced.

CASE 1:

The original dataset was given as an input to the genetic algorithm (GA) to pre-process, and the data subset with selected features, from the GA, was then given as an input to the filter and the previously mentioned embedded feature selection algorithms (FSA). The latter pre-processing yielded nine (9) subsets which were then used as input by the random forest classifier for prediction. Figure 1 below shows the flow diagram used in this case.

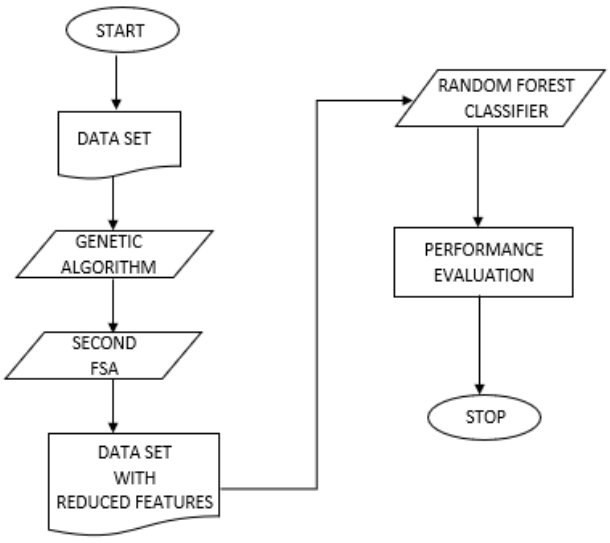

FIGURE 1. Flow diagram of case 1 hybridization predictive model.

CASE 2:

The original dataset was passed as input to the nine FSAs, which led to nine data subsets with selected features based

on the methods of each FSA. The nine resultant subsets were given as input to the GA. The output (nine subsets) obtained from the GA were then passed to the random forest classifier for prediction. Figure 2 below shows the flow diagram used in this case.

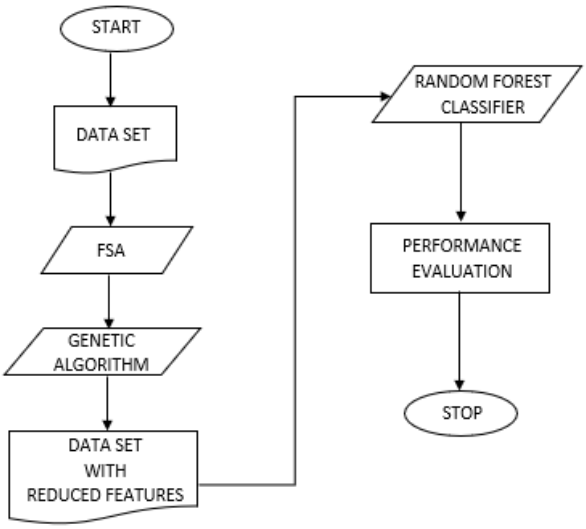

FIGURE 2. Flow diagram of case 2 hybridization predictive model.

CASE 3:

In this case, a union of features was formed from subsets derived in cases 1 and 2. This process also yielded nine (9) subsets which were then used as input by the random forest classifier for prediction. Figure 3 shows the flow diagram used in case 3.

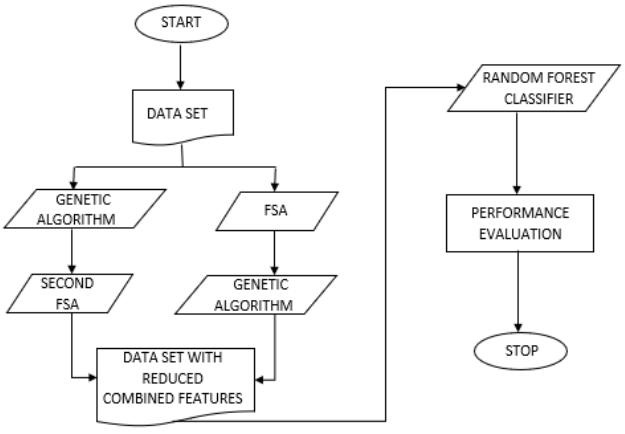

FIGURE 3. Flow diagram of case 3 hybridization predictive model.

IV. PERFORMANCE EVALUATION

In order to assess the effectiveness of our proposed models, we will employ various performance metrics such as Accuracy, Precision, Recall, and F1 score. As indicated in Table 1, the genetic algorithm (GA) was initially used as a feature selection method, without hybridization, in random forest (RF) predictive model, this model was observed to be high-performing with a performance accuracy of 98.23. The only CASE 1 hybridized predictive model that out-performed the earlier model mentioned was the GA + Fisher_Score predictive model, with 0.89 increase in accuracy. GA + Variance model also have a high performance of 97.37, while GA + Correlation [0.4] and GA + L2 (Ridge) had the same accuracy value of 96.49. Correlation [0.4] + GA and L1 (Linear) + GA predictive models had 100% precision for malignant tumors (1) and 100% Recall for benign tumors (0).

TABLE 1. Predictive experimental results obtained from case 1 (0-benign, 1- malignant tumor).

S/N	FSA MODEL	ACCURACY		PRECISION	RECALL	F1
1	GA	**98.23**	**0**	**99.00**	**99.00**	**99.00**
			1	**98.00**	**98.00**	**98.00**
	GA + Correlation [0.5]	93.86	0	95.00	96.00	95.00
			1	92.00	90.00	91.00
	GA + Correlation [0.4]	96.49	0	97.00	97.00	97.00
			1	95.00	95.00	95.00
2	GA + Variance	97.37	0	96.00	100.00	98.00
			1	100.00	93.00	96.00
3	GA + L1 (Linear)	92.11	0	88.00	98.00	93.00
			1	98.00	85.00	91.00
4	GA + L1 (Lasso)	94.74	0	93.00	99.00	96.00
			1	98.00	89.00	93.00
5	GA + L2 (Ridge)	96.49	0	97.00	97.00	97.00
			1	95.00	95.00	95.00
6	GA + Tree-based	94.74	0	96.00	96.00	96.00
			1	92.00	92.00	92.00
7	GA + MIG	94.74	0	96.00	96.00	96.00
			1	93.00	93.00	93.00
8	GA + Fisher_Score	**99.12**	0	**100.00**	**99.00**	**99.00**
			1	**97.00**	**100.00**	**99.00**
9	GA + Chi-test	95.61	0	93.00	100.00	96.00
			1	100.00	90.00	96.00

For the CASE 2 hybridization methodology, the Tree-based + GA predictive model was seen to have out-performed all other predictive models, with a performance accuracy of 97.37, as shown in Table 2. It was closely followed by Correlation [0.4] + GA, L1 (Linear) + GA, and MIG + GA predictive models, they had same accuracy result of 96.49 which was about 0.88 lesser than the best performing model in this category. Correlation [0.4] + GA and L1 (Linear) + GA predictive models had 100% precision for malignant tumors (1), 100% Recall for benign tumors (0).

In Table 3, the GA ∪ L1 (Lasso) predictive model had the best performance for the CASE 3 methodology with an accuracy value of 98.25 with a100% precision for malignant

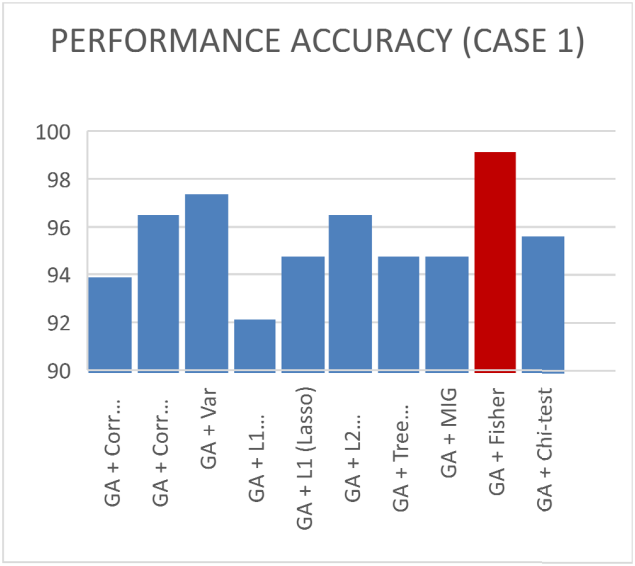

FIGURE 4. Graphical illustration of CASE 1 performance accuracy.

TABLE 2. Predictive experimental results obtained from case 2 (0-benign, 1- malignant tumor).

S/N	FSA MODEL	ACCURACY		PRECISION	RECALL	F1
1	Correlation [0.5] + GA	93.86	0	88.00	100.00	94.00
			1	100.00	82.00	90.00
	Correlation [0.4] + GA	96.49	0	96.00	100.00	98.00
			1	100.00	93.00	96.00
2	Variance + GA	92.11	0	93.00	94.00	93.00
			1	91.00	90.00	91.00
3	L1(Linear) + GA	96.49	0	94.00	100.00	97.00
			1	100.00	91.00	95.00
4	L1 (Lasso) + GA	92.11	0	90.00	99.00	94.00
			1	97.00	81.00	89.00
5	L2 (Ridge) + GA	92.98	0	92.00	97.00	94.00
			1	95.00	86.00	90.00
6	Tree-based + GA	**97.37**	0	**97.00**	**99.00**	**98.00**
			1	**97.00**	**95.00**	**96.00**
7	MIG + GA	96.49	0	97.00	97.00	97.00
			1	96.00	96.00	96.00
8	Fisher_Score + GA	92.11	0	96.00	92.00	94.00
			1	86.00	92.00	89.00
9	Chi-test + GA	93.86	0	93.00	97.00	95.00
			1	95.00	89.00	92.00

tumors (1), 100% Recall for benign tumors (0) and 99% f1 score for benign tumors (0). This performance was closely followed by a 97.37 accuracy from GA ∪ Correlation [0.5], GA ∪ Correlation [0.4], GA ∪ Variance, and GA ∪ Chi-test predictive models.

TABLE 3. Predictive experimental results obtained from case 3 (0-benign, 1- malignant tumor).

S/N	FSA MODEL	ACCURACY		PRECISION	RECALL	F1
1	GA ∪ Correlation [0.5]	97.37	0	97.00	99.00	98.00
			1	98.00	96.00	97.00
	GA ∪ Correlation [0.4]	97.37	0	99.00	97.00	98.00
			1	95.00	98.00	96.00
2	GA ∪ Variance	97.37	0	99.00	97.00	98.00
			1	95.00	98.00	97.00
3	GA ∪ L1 (Linear)	95.61	0	94.00	99.00	96.00
			1	98.00	91.00	95.00
4	GA ∪ L1 (Lasso)	**98.25**	0	**97.00**	**100.00**	**99.00**
			1	**100.00**	**95.00**	**97.00**
5	GA ∪ L2 (Ridge)	96.49	0	95.00	100.00	97.00
			1	100.00	91.00	95.00
6	GA ∪ Tree-based	94.74	0	93.00	99.00	96.00
			1	98.00	89.00	93.00
7	GA ∪ MIG	96.49	0	96.00	99.00	97.00
			1	98.00	93.00	95.00
8	GA ∪ Fisher_Score	96.49	0	95.00	98.00	97.00
			1	98.00	94.00	96.00
9	GA ∪ Chi-test	97.37	0	97.00	99.00	98.00
			1	97.00	95.00	96.00

V. RESULTS

From the performance evaluation, we can observe that some models had the same accuracy in two of three methodologies (cases) considered. For instance, GA + L2 (Ridge) and GA ∪ L2 (Ridge) predictive models were observed to have had the same performance accuracy of 96.49. Also, GA + Variance and GA ∪ Variance predictive models were observed to have had the same performance accuracy of 97.37.

GA + MIG had an accuracy of 94.74, while MIG + GA and GA ∪ MIG had a higher value of 96.49. Similarly, GA + Tree-based and GA ∪ Tree-based had same accuracy values of 94.74, while Tree-based + GA had a higher value of 97.37.

We also observed that some predictive models had different accuracies in all three methodologies (cases) considered. For instance, GA + Chi-test had an accuracy of 95.61 and Chi-test + GA had 93.86 while GA ∪ Chi-test obtained 97.37.

Also, GA + L1 (Linear) had an accuracy 92.11, and L1 (Linear) + GA obtained 96.49 and GA ∪ L1 (Linear) achieved 95.61 accuracy.

Lastly, GA + L1 (Lasso) achieved 94.74 accuracy, and L1 (Lasso) + GA had 92.11, while GA ∪ L1 (Lasso) obtained 98.25.

From the several instances aforementioned, we can see that the placement order does have an impact on the performance of a hybridized predictive model, therefore cross checking the placement order when hybridization is done

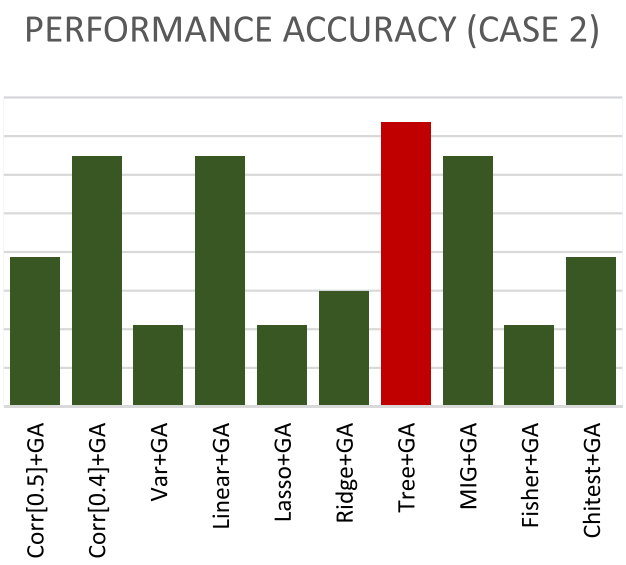

FIGURE 5. Graphical illustration of CASE 2 performance accuracy.

will be a good practice. In this research, the proposed combined methodology of Cases 1 and 2, that is Case 3, was observed to have produced predictive models with higher accuracies in comparison to those in Cases 1 and 2. Although, not all combined predictive models had higher accuracies, as the overall best performing predictive model was found in Case 1, which is the GA + Fisher_Score with an accuracy of 99.12.

VI. DISCUSSION

Hybridized feature selection method is an important factor in predictive model, as it is focused on achieving optimal feature subset from large dataset, by combining the strength of filter, wrapper, and embedded methods to take out features with little to no relevance in the dataset. Thereby, improving accuracy, efficiency and robustness of predictive models even when applied to large datasets. However, not taking into cognizance the order of hybridization can have an impact on the hybridized model performance.

In this paper, we conducted several hybridizations of genetic algorithm with filter and embedded feature selection methods, in the data pre-processing phase of Random Forest predictive model, with the aim of improving its performance. Specific focus was placed on the FSA placement order considered in each predictive model, as we considered three possible hybridization case studies; initially order, reversed order and combined order.

From our performance evaluation, the predictive model containing GA + Fisher_Score hybridized feature selection was observed to outperform all other hybridized feature selection methods considered in the research, including its reversed model; Fisher_Score + GA. We observed that Feature selection algorithms select a set of features based on their unique criteria. Therefore, when the first FSA selects a feature subset from the initial dataset, based on its criteria, the second FSA then further refines that feature subset. The choice of the algorithm applied first can have a significant impact on the final feature subset selected. If the first and second FSAs have similar criteria for selecting features, then the order may not matter as much. However, if there is a significantly different in their criteria, then the order can have a more significant impact on the final feature subset selected.

The criteria are factors or rules used by FSAs in selecting the best feature subset for the machine learning predictive model. The criteria may vary depending on the FSA, often they are based on measures such as relevance, redundancy, and stability of the features [18], [19], [20], [21], [22]. For instance, one FSA may use a relevance criterion to select features that are most informative and correlated with the target variable. Another FSA may use a redundancy criterion to remove features that are highly correlated with each other to avoid overfitting. A stability criterion may be used to select features that are consistently selected across multiple iterations or data subsamples.

Hence, not only is the hybridization of feature selection models important in prediction, but it is also important to carefully consider the criteria of the FSAs being hybridized and their order of application to ensure optimal performance of the machine learning predictive model. It may be necessary to try multiple orders and compare their performance to determine the best order for the specific problem at hand.

The hybridization of Genetic Algorithm with Fisher_Score for feature selection was seen to provide a powerful approach to prevent overfitting by leveraging the complementary strengths of both methods, that is, the exploration capabilities of GA and the discriminative power assessment of Fisher_Score to find feature subsets that are both relevant and non-redundant. It reduces dimensionality, as we can identify the most informative and discriminative features. This dimensionality reduction helps to prevent overfitting by removing irrelevant or redundant features that may introduce noise or bias into the learning process. Also, generalization ability is enhanced based on the discriminative power of the Fisher_Score, as selected features are more likely to capture the underlying patterns and relationships in the data, rather than spurious correlations or noise. This focus on relevant features can lead to more robust and accurate models that are less prone to overfitting.

Genetic Algorithms (GA) are capable of exploring large solution spaces efficiently. By incorporating Fisher_Score as a fitness function or as a guiding criterion in the GA process, the search is guided towards more promising regions of the search space that contain relevant features. This can accelerate the convergence of the GA and improve the overall efficiency of the feature selection process.

Anomalous or outlier data points can have a significant impact on the performance and generalization of machine learning models. However, by incorporating Fisher_Score

to assess the discriminative power of features, the model assigns higher importance to features that contribute more to the separation of classes or target variables. This helps in excluding features that may be sensitive to anomalies or contain excessive noise, resulting in a more robust and reliable model.

The computational complexity of a GA + Fisher_Score model depends on various factors, such as the size of the dataset, the number of features, and the specific implementation details. **Genetic Algorithm (GA):** The computational complexity of a genetic algorithm is primarily determined by the number of iterations and the population size. Each iteration involves evaluating the fitness of each individual in the population, performing selection, crossover, and mutation operations. The complexity is typically O (I * P * F), where I is the number of iterations, P is the population size, and F is the time complexity of evaluating the fitness function.

Fisher_Score: The computational complexity of calculating the Fisher_Score depends on the number of classes and the number of features. For a dataset with N samples and M features, the complexity is generally O (N * M).

Feature Selection: The complexity of feature selection using the Fisher_Score depends on the specific method and algorithm employed. For a greedy search approach, such as sequentially adding or removing features based on their Fisher_Score, the complexity is typically O (M^2) or O (M * N) for N iterations. For more sophisticated techniques like branch and bound, the complexity can be higher. Combining all these components, the overall computational complexity of the GA + Fisher_Score model will depend on the interplay between the GA iterations, population size, feature dimensionality, and the specific feature selection algorithm used.

It is also important to note that these complexities are approximate and may vary depending on the specific implementation details and optimizations applied. Additionally, the computational complexity does not account for the time required for data preprocessing, model training, or other auxiliary tasks.

We compared the performance accuracy of our proposed model with those found in recent literature, to verify the robustness and validity of the proposed GA+ Fisher_Score model. Table 4 below shows this comparison. We observed that the performance accuracy (99.00) in [24] was quite similar with that of our proposed model (99.12). However, it is important to note that Fisher_Score is particularly effective when classes are well-separated, as it captures the discriminative power of features and ranks them based on their ability to discriminate between classes. Fisher_Score emphasizes informative features, making it ideal for identifying relevant features for classification tasks. It is computationally efficient, especially for high-dimensional datasets, and is less prone to overfitting, providing more reliable feature rankings. However, the choice of feature selection method depends on the dataset and task requirements, so it is recommended to experiment and evaluate performance for the specific scenario.

TABLE 4. Performance comparison of related works with our proposed model, using Wisconsin dataset.

AUTHOR	CONTRIBUTION	ACCURACY
Farid, *et al.* (2020) [23]	Hybrid F.S. optimization of GA on AdaBoost Stacked-Hybrid Classification -with SVM model	98.25
Vutakuri and Maheswari (2020) [24]	Combined mutual information and genetic algorithm {MIGA)	99.00
Abd-elnaby, *et al.* (2022) [25]	Hybrid mutual information, LASSO and genetic algorithm (MI-LASSO-GA)	95.00
Ali and Saeed (2023) [26]	Hybrid F.S information gain ratio and genetic algorithm IGR-GA-with RF model.	93.81
Our Proposed GA + Fisher_Score Model	**Hybrid F.S. GA and Fisher_ Score with RF Model**	**99.12**

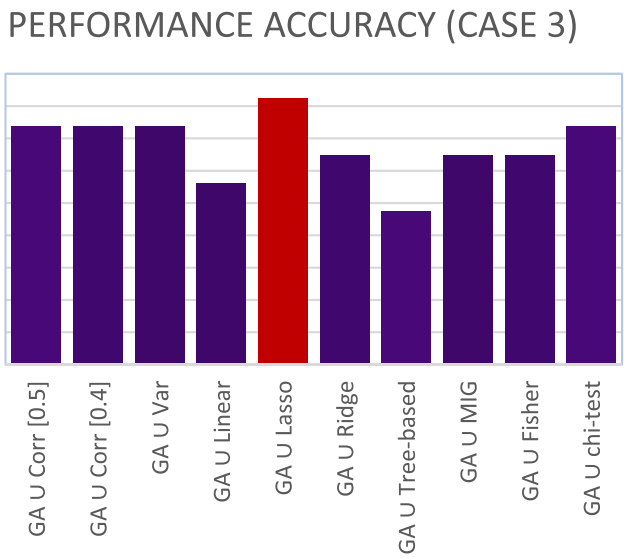

FIGURE 6. Graphical illustration of CASE 2 performance accuracy.

VII. CONCLUSION AND FUTURE WORK

In conclusion, this paper explored three classes of feature selection techniques: Filter methods, Wrapper methods, and Embedded methods. Filter methods evaluate feature relevance based on correlation with the dependent variable, while Wrapper methods use model training to measure the usefulness of feature subsets. Embedded methods enhance the objective function during the learning process. The study used the Wisconsin Hospitals Madison Breast Cancer Database, which consisted of 569 samples and 32 features. The Random Forest classifier was employed to predict benign and malignant tumors. The research compared different hybridization methodologies and evaluated their performance

using accuracy as the metric. In Case 1, the hybrid model GA + Fisher_Score achieved the highest accuracy of 99.12%. In Case 2, the Variance + GA hybrid model performed the best with an accuracy of 97.37%. Case 3, a combination of features from Cases 1 and 2, yielded the GA ∪ L1 (Lasso) hybrid model with the highest accuracy of 98.25%. The placement order of the feature selection algorithms was found to have an impact on the final feature subset selected.

The study highlighted the importance of hybridized feature selection methods in improving the performance of predictive models. It emphasized the need to consider the criteria and order of the feature selection algorithms being hybridized. The GA + Fisher_Score hybridization was particularly effective in preventing overfitting and improving generalization. However, the choice of feature selection method should be based on the specific dataset and task requirements. Computational complexities were discussed, and comparisons with previous literature validated the proposed GA + Fisher_Score model's robustness.

The shortcomings of the GA + Fisher_Score algorithm include sensitivity to dataset characteristics, limited scalability for large datasets, the need for careful parameter tuning, and lack of interpretability. Despite these limitations, the hybridized feature selection approach has the potential to enhance the performance of predictive models and improve accuracy, efficiency, and robustness when applied to large datasets.

Overall, the research demonstrated the benefits of hybridized feature selection techniques and provided insights into the order and criteria considerations for optimal performance. The findings contribute to enhancing the accuracy, efficiency, and robustness of predictive models in the context of large datasets.

In the future, we plan to work on including ensemble learning in the predictive methodology, as well as, use alternative datasets to optimize the model and improve its performance.

REFERENCES

[1] *Cancer Facts and Figures 2022, Atlanta: American Cancer Society*, American Cancer Society, Atlanta, GA, USA, 2022.

[2] (2022). *Breast Cancer Facts and Statistics*. [Online]. Available: https://www.breastcancer.org/facts-statistics

[3] A. R. Vaka, B. Soni, and S. Reddy, "Breast cancer detection by leveraging machine learning," *ICT Exp.*, vol. 6, no. 4, pp. 320–324, Dec. 2020.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[5] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018.

[6] K.-H. Chen, K.-J. Wang, A. M. Adrian, K.-M. Wang, and N.-C. Teng, "Diagnosis of brain metastases from lung cancer using a modified electromagnetism like mechanism algorithm," *J. Med. Syst.*, vol. 40, no. 1, pp. 1–14, Jan. 2016.

[7] J. Ayoola and T. Ogunfunmi, "A comparative analysis of regression algorithms with genetic algorithm in the prediction of breast cancer tumors," in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, Sep. 2022, pp. 143–149.

[8] Y. Sun, C. F. Babbs, and E. J. Delp, "A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive sequential floating search vs. genetic algorithm," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jan. 2005, pp. 6532–6535.

[9] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, "Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information," in *Proc. Int. Conf. Interact. Technol. Games (ITAG)*, Oct. 2016, pp. 70–76.

[10] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, "A comparative study for breast cancer prediction using machine learning and feature selection," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 1049–1055.

[11] K. Nouira, Z. Maalej, F. B. Rejab, L. Ouerfelly, and A. Ferchichi, "Analysis of breast cancer data: A comparative study on different feature selection techniques," in *Proc. Int. Multi-Conf., 'Org. Knowl. Adv. Technologie' (OCTA)*, Feb. 2020, pp. 1–11.

[12] J. Suto, S. Oniga, and P. P. Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," in *Proc. 6th Int. Conf. Comput. Commun. Control (ICCCC)*, May 2016, pp. 124–129.

[13] Q. Wu, Z. Ma, J. Fan, G. Xu, and Y. Shen, "A feature selection method based on hybrid improved binary quantum particle swarm optimization," *IEEE Access*, vol. 7, pp. 80588–80601, 2019.

[14] X. Zhou, Q. Wang, R. Zhang, and C. Yang, "A hybrid feature selection method for production condition recognition in froth flotation with noisy labels," *Minerals Eng.*, vol. 153, Jul. 2020, Art. no. 106201.

[15] A. Kawamura and B. Chakraborty, "A hybrid approach for optimal feature subset selection with evolutionary algorithms," in *Proc. IEEE 8th Int. Conf. Awareness Sci. Technol. (iCAST)*, Nov. 2017, pp. 564–568.

[16] D. Jain and V. Singh, "Diagnosis of breast cancer and diabetes using hybrid feature selection method," in *Proc. 5th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, Dec. 2018, pp. 64–69.

[17] B. Sahmadi and D. Boughaci, "Hybrid genetic algorithm with SVM for medical data classification," in *Proc. Int. Conf. Appl. Smart Syst. (ICASS)*, Nov. 2018, pp. 1–6.

[18] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[20] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2019, pp. 442–452.

[21] P. Shen, X. Ding, W. Ren, and S. Liu, "A stable feature selection method based on relevancy and redundancy," *J. Phys., Conf.*, vol. 1732, no. 1, Jan. 2021, Art. no. 012023.

[22] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, 2019.

[23] A. A. Farid, G. Selim, and H. Khater, "A composite hybrid feature selection learning-based optimization of genetic algorithm for breast cancer detection," Tech. Rep., 2020, doi: 10.20944/preprints202003.0298.v1.

[24] N. Vutakuri and A. U. Maheswari, "Breast cancer diagnosis using a Minkowski distance method based on mutual information and genetic algorithm," *Int. J. Adv. Intell. Paradigms*, vol. 16, nos. 3–4), pp. 414–433, 2020.

[25] M. Abd-elnaby, M. Alfonse, and M. Roushdy, "A hybrid mutual information-LASSO-genetic algorithm selection approach for classifying breast cancer," in *Digital Transformation Technology*. Singapore: Springer, 2020, pp. 547–560.

[26] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data," *Processes*, vol. 11, no. 2, p. 562, 2023.

JOYCE A. AYOOLA received the B.S. and M.S. degrees in computer science from Landmark University, Nigeria, in 2014 and 2020, respectively. She is currently pursuing the Ph.D. degree with the Information and Machine Research Group, Department of Electrical and Computer Engineering, Santa Clara University, Santa Clara, CA, USA. Her research interests include artificial intelligence and machine learning, with a passion for health informatics.

TOKUNBO OGUNFUNMI (Senior Member, IEEE) received the B.S. degree (Hons.) from Obafemi Awolowo University (formerly University of Ife), Ife, Nigeria, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA.

He was a Visiting Professor with The University of Texas at Austin, Austin, and Stanford University, and as a Carnegie Foundation Visiting Professor. From 2010 to 2014, he was the Associate Dean of the Research and Faculty Development, School of Engineering, Santa Clara University (SCU), Santa Clara, CA, USA. At SCU, he teaches a variety of courses in circuits, systems, signal processing, and related areas. He is currently a Professor of electrical and computer engineering and the Director of the Information Processing and Machine Learning Research Laboratory, SCU. His current research interests include digital and adaptive signal processing and applications, machine learning, deep learning, speech and multimedia (audio, video) compression, and nonlinear signal processing. He has published over 200 refereed journal and conference papers in these areas.

Dr. Ogunfunmi is currently serving on the editorial board of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the *Circuits, Systems, and Signal Processing* (CSSP) journal. He has been involved with several IEEE conference committees as a member of the organizing and technical committees. He served as the General Chair for the 2018 IEEE Workshop on Signal Processing Systems (SiPS 2018) and the Technical Program Co-Chair for the 2019 IEEE International Symposium on Circuits and Systems (ISCAS 2019). He also served as a Lead Guest Editor for the CSSP journal Special Issue on "Algorithms and Architectures for Machine Learning Based Speech Processing," published in August 2019, and the *Journal of Signal Processing Systems* (JSPS) Special Issue on 2018 IEEE Workshop on Signal Processing Systems (SiPS). From 2013 to 2014, he served as a Distinguished Lecturer for the IEEE Circuits and Systems Society.