International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Multivariate Analysis on Personalized Cancer Data using a Hybrid Classification Model using Voting Classifier

Ashok Reddy Kandula[1], Dr. R. Sathya[2] and Dr. S. Narayana[3]

Submitted: 01/11/2022 **Revised:** 23/12/2022 **Accepted:** 07/01/2023

Abstract: Cancer was found to be a significant burdening type of problem in the medical system. An accurate diagnosis is considered to be a challenging task for physicians. Modern Artificial Intelligence (AI) research proves that cancers are easily identifiable and early detectable and can be diagnosed by classification with the help of gene mutations that occur in the cells. This paper presented a hybrid classification model using an ensemble machine learning technique to classify the type of cancer class from the given medical dataset. The proposed hybrid voting classifier utilizes guided WOA (Whale Optimization Algorithm) that aggregates different classification models' prediction results to choose the most voted class. This helps to increase the chance that individual classifiers, e.g., Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF), show significant discrepancies. The results were found to be better compared to the existing work.

Keywords: Artificial Intelligence, Medical data analysis, gene mutation, ensemble, Logistic Regression, Support Vector Machine, Random Forest.

1. Introduction

Machine Learning (ML) techniques significantly handle massive amounts of data, even when the data is heterogeneous, missing, or inconsistent. The medical analysis is considered one of the major applications of machine learning. While moving on to the medical problems, cancer was considered problematic when it was found at the last stage. Luckily with the help of machine learning, medical practitioners could identify cancer diseases with the help of advanced models and techniques that predict the cancer disease and type of cancer by training with enormous data samples. The perfect model with data processing and classification was considered a next-generation information technology for cancer diagnosis, prognosis, and prediction [1].

Various mandatory processes are performed in the machine learning techniques to obtain the best results from the given dataset, from learning the data and classifying the data to predicting the outcome results. In our previous work [2], a complete data visualization technique was performed to select the best and most important feature from the dataset and identify the feature importance and stability towards classification using univariate analysis on each feature. In addition, the class distributions are found to achieve imbalanced dataset problems effectively. After the data visualization, several machine learning techniques like Naive Bayes, K-Nearest Neighbours, and Logistic Regression with class balancing and without class balancing [3] were applied to select the best methodology for the given dataset. Therefore, this methodology proposes a hybrid classification model using an ensemble machine learning technique to classify the cancer types.

To get the best results, the work utilized advanced machine learning algorithms like Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) as base learners under a hard voting system to classify the appropriate cancer class. The reason for selecting LR as it yields good performance among the other ML models to predict the risk of major cancer diseases with improper data. In the existing work, the appropriate results are discussed, including both class balancing and class unbalancing. While the reason for choosing SVM as the data model with maximization is support in the process of separating margins that is a vector which is combined to form Support Vector Machine (SVM) technique learning, which was considered a powerful classification tool that has been utilized towards cancer class classification or to be specific subtyping. To perform the ensemble learning, the

[1]*Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu-608002, India. E-mail: ashokreddy.gec@gmail.com*
[2]*Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu-608002, India. E-mail: sathya.aucse@gmail.com*
[3]*Professor, Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru-521356, India. E-mail: satyala1976@gmail.com*

work utilizes another machine learning model, Random Forest (RF), which was also considered to have an ensemble technique while making decisions[17] that solves classification and regression problems. Addressing the machine learning problem, i.e., to classify with accuracy, ensemble learning is utilized in which a machine learning scheme boosts the accuracy by integrating multiple models like Logistic Regression, Support Vector Machine, and Random Forest. Multiple classification models contribute to ensemble learning to obtain more accurate results than single classifiers. While performing multiple classifier integrations, the model tends to decrease variance, mostly in the case of unstable classifiers, which produce reliable results. Following the scenario, a voting classification model was designed to assign a label to unlabelled samples. The commonly used voting approach is the majority voting classifier that assigns the class labels with a maximum number of votes from various other classifiers to each separate unlabeled data. The majority voting classifier was popular as it holds its simplicity and effectiveness. The advanced voting classifier utilized here is a soft voting classifier that uses probabilities of the class labels from the classifier.

To find the best possible solution for any particular problem, Optimization is the technique to find all available solutions. Among the techniques, one such powerful method to solve several applications in medical problems is Meta-heuristic algorithms. These algorithms are highly inspired by physical algorithms and their logical behaviors that are found in nature. The existing solutions found that these optimization techniques can be performed with less computational effort and within a reasonable time.

The proposed hybrid voting classifier utilizes guided WOA (Whale Optimization Algorithm) that aggregates different classification models' prediction results to choose the most voted class. This helps to increase the chance that individual classifiers, e.g., Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF), show significant discrepancies.

However, providing an accurate diagnosis was a challenging task for medical practitioners. To provide appropriate treatment, an automatic prediction system to diagnose the cancer patient is mandatory at this period. The work addresses the following objectives(a) To address the problems in the existing system with our solutions to diagnose the patient foremost (b) to provide specialized machine learning technology and expertise to predict cancer class with given data. (c) To provide an accurate prediction of cancer with accuracy measures and probability values.

2. Literature Review

The work has undergone detailed study covering the important medical data analysis in machine learning technology, following to study of major machine learning algorithms. [5]presented an optimization based on a parallelization algorithm and hierarchical selection based on maximizing the relevance sum and distances to solve high dimensionality issues. This model incorporated parallel optimization and K- means with pruning on the candidate model with a hierarchical selection model. The paper combined each basic model's prediction results by a majority voting technique that employed a divide-and-conquer-based approach that solved computation issues.

Based on feature selection and ensemble [6] presented an intrusion detection model where heuristic algorithm CFS-BA been utilized for dimensionality reductions that selects feature correlation with optimal subset process with ensemble approach combining Random Forest (RF), C4.5, and Forest through Penalizing Attributes algorithms which processed by voting classification.

In [7], built a classifier based on data streams, namely the Parsimonious Ensemble (pENsemble) approach that incorporated dynamic online feature selection that allows selection and de-selection of input features. [8] addressed the uncharacterized protein sequence with the help of an ensemble predictor using a voting system, in [9] utilized sentiment analysis to analyze the attitude, emotions, and opinions of different people in the micro-blogging platform Twitter to collect views related to products, politics and trends. It used a voting classifier (Logistic Regression with SGD classifier to classify tweets for emotion recognition.

To address heart disease in humans, [10] presented a majority voting ensemble to train and predict from real-life data samples of healthy and ill people. [11] presented an ensemble classification for the Wisconsin Breast Cancer Dataset (WBCD) that utilized Logistic regression learning, Support Vector included with stochastic gradient descent optimization and multilayer perceptron network. The paper also evaluated the performance based on soft and hard voting with the help of probabilities to find averages, product, maximum and minimum probabilities with a voting classifier.

[12] addressed the prevention steps in Corona Virus with similar diseases like pneumonia and processing chest CT images[13] to diagnose COVID-19. To address this issue, this paper utilized machine learning techniques with two optimization algorithms to perform feature selection and classification for COVID-19. In [14] addressed bug reports which are an inescapable part of the software product frameworks with studying ensemble classification methods for new incoming bugs. The paper incorporated five types of ensemble techniques Boosting, Bagging,

average voting, Majority voting, and stacking, using a few machine learning classifiers as base classifiers.

Generally, ensemble classifiers either achieve dataset features or exploit data sample space. Some classifiers like RF exploits input dataset attribute while bagging exploits data sample spaces. [15] presented an approach that first reduces dataset attribute space by preferring to pick only significant features from the dataset, leading to maximizing the ensemble classification performance.

In some places, the base classifiers could perform differently; thus, in [16] assigned, different weights to increase the performance of the classification. The work presented a novel weighted Majority Voting Ensemble (WMVE) technique that adjusts each classifier on its contribution towards class decision-making.

The paper is organized as part I is the Introduction. In contrast, part II is a literature review, Part III is materials and methodology, Part IV is Results and Discussions, and finally provided conclusion in part V.

3. Materials and Methodology

Use This research work is highly based on improving the medical dataset's performance and accuracy that classifies cancer types among several cancer types. The work proposed a hybrid classification model using an ensemble machine learning approach for multi-class classification of cancer types by differentiating between the names Class 1 to class 9. The existing work of the paper holds the proper data collection and visualization techniques applied to it. At the same time, the proposed model proposed a hybrid model that holds well-defined collective machine learning models together applied in it.

3.1 Existing System

The Existing system of this personalized medical dataset proceeded with only one machine-learning algorithm at a time. Our previous paper implemented the work in the existing system by training the data model with each machine-learning algorithm separately to get accurate results. The used machine learning models are Naïve Bayes, K-Nearest Neighbors, and logistic regression with class balancing and logistic regression without class balancing. The previous work provided the best approach to selecting the appropriate machine learning algorithm by implementing several algorithms separately with the same data model. Among that, Logistic regression without class balancing was found to be effective by giving less error rate for the model. In the previous works, data collecting and visualizing were performed and elaborated well by providing proven results.

3.1.1 Data Collection

The most significant task for the machine learning approach is collecting the appropriate data that is medical-related terms. Though the medical data is unique and hard to collect, the Kaggle platform has some publicly available datasets which could be utilized for study purposes. Kaggle is the platform provided by Google LLC subsidiary that gathers data for research processes and allows researchers to search and exchange datasets online. The dataset utilized for this study is a medical dataset containing training data that holds three major features: gene feature, variant feature, and text feature.

3.1.2 Data Preprocessing

Data pre-processing is an important phase before applying any machine learning models. Data pre-processing changes the data into a usable, efficient format that can be inputted into the algorithms for applying machine learning techniques. The first data pre-processing step applied was label encoding which changes the label-given format into the machine-learnable format as a binary encoding. That is followed by text pre-processing, which includes removing stop words, converting special characters with spaces, changing multiple spaces into a single space, and converting the upper case into lower case.

3.1.3 Data Visualization

To understand the complete structure of the given data, the dataset's visualization has been carried out with multiple visualizations and statistical techniques combined with univariate analysis of each feature shown in previous work. The dataset contains 3321 data points which it further separated into train data, cross-validation (CV) data, and testing data under the ratio 64:20:16, where the data points are separated into a train as 2124, CV as 532, test as 665 data points respectively. While the class points from class 1 to class 9 represent types of cancers associated with each data point. Class 1 to class 9, class 7, class 4, class 1, and class 2 were found to dominate the other classes.

3.2 Proposed Model

The hybrid system proposed in this work uses three popular best-performing machine learning models, Logistic Regression (LR), Support Vector Machines (SVMs), and Random Forest (RF), for classifying the data to predict the class labels.

3.2.1 Model Architecture

The model architecture follows through several stages data cleaning, pre-processing techniques, and hyperparameter tuning are performed with the stacking of three classifiers; later soft voting classifier is applied.

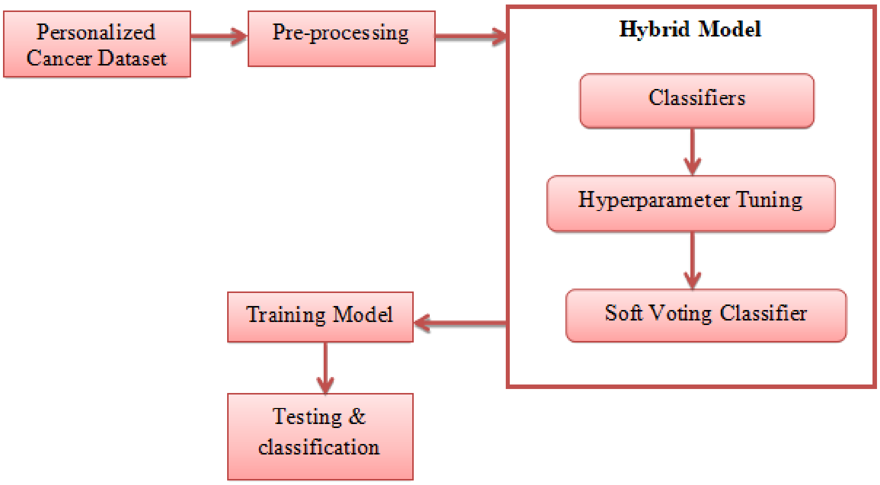


Fig 1. Model Architecture

3.2.2. Logistic Regression

Logistic Regression (LR) is one of the techniques borrowed by machine learning from the field of statistics. The Logistic regression was named for the use of its core function, which is also called as sigmoid function developed by statisticians [4]. The logistic regression was an S-Shaped curve that takes whatever the real-valued number to map into values of 0 and 1, which is still exactly never limited to those values provided in a mathematical equation (1).

$$CLR = p(Y = 1) = \frac{e^t}{1 + e^t}$$

(1)

Where e is denoted for the constant that holds Euler's number '2.718' while t can be any linear combination of predictors like b0 + b1x, the (Y=1) indicates the probabilities.

3.2.3 Support Vector Machine

Support Vector Machine (SVM) was considered a supervised machine learning algorithm that solves both classification and regression issues. It is most commonly used in classification problems. An SVM classifier is one of the high-performing techniques among machine learning algorithms that work based on dividing data into different regions [17]. The SVM tends to find the maximum margin that always separates the dataset into two separate groups and to determine under which category any upcoming data falls. Most of the researchers prefer SVMs as it tries to use less computing power to produce significant accuracy. SVMs also perform exceptionally well on any sized dataset, from smaller to larger, efficiently handling high-dimensional spaces and memory. The advanced model of SVM, Radial Basis Function (RBF), which can also be mentioned as Gaussian Kernel as mentioned in equation (2), has been used to result in a complex decision boundary.

$$C_{SVM}(X_i, X_j) = \exp\left(-\gamma |X_i - X_j|^2\right).$$

(2)

3.2.4 Random Forest

Random Forest (RF) is a supervised machine learning classifier that is simple, diversified, and versatile. RF solves both classification and regression problems. It builds huge forest-like structures, the ensembles of decision tree models that achieve better prediction results [18]. In classification, decision trees generally work individually to predict a class's outcome, in which the final prediction will be the class with the majority votes.

$$C_{RF} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \mu|$$

(3)

Where$y_i$label, for instance, N indicates several instances, and μ indicates the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$.

3.2.5 Ensemble learners

Ensemble learners are always used to improve the performance of the model. In general ensemble, the technique combines the prediction results of two or more classifiers to create a model that tends to give more accurate results. The base logic behind this technique is similar to our daily life activity, which is obtaining experts' opinions before taking a final decision. The last ensemble-based machine learning was the method that reduced risk in decision-making. A major example of an ensemble approach is voting classifiers, where the end decision or classification is drawn based on the majority votes counted from all the algorithms taken. The ensemble approach has been used in vast applications like spam detection, face recognition, text categorization, character recognition, etc. Wherever machine learning techniques could be used, an ensemble could be applied at that place. Thus voting classifier will be utilized in this paper as an ensemble technique

3.2.6 Whale Optimization Algorithm

The whale Optimization algorithm (WOA) is the inspiration obtained from the behaving capabilities of the whales, where the whales use bubbles to trap their prey which tends to force them to reach the surface in a spiral shape. In the mathematical computations, the first step of the optimizer is based on the equation below:

$$\vec{G}(t+1) = \vec{G}*(t) - \vec{A}.\vec{D}, \vec{D} = \left|\vec{C}.\vec{G}*(t) - \vec{G}(t)\right| \quad (4)$$

In the above equation, where $\vec{G}(t)$ represents a solution factor for the t iteration, the vector $\vec{G}*(t)$ indicates the position of the prey. The "." Indicates the pairwise multiplication, and the $\vec{G}(t1)$ indicates the updated position for the above solution. The other two vectors, $\vec{A}$ and $\vec{C}$, were updated over the iteration through $\vec{A} = 2\vec{a}.\vec{r_1} - \vec{a}$ and $\vec{C} = 2.\vec{r_2}$ for the vector $\vec{a}$ linearly changes its values from 2 to 0, and $\vec{r_1}$ and $\vec{r_2}$ were considered as random values picked between [0,1].

In the process, the shrinking encircling was considered the second mechanism, including the decreasing behavior of the vector values $\vec{a}$ and $\vec{A}$, followed by the spiral procedure by updating the positions, which was mathematically written as in the equation.

$$\vec{G}(t+1) = \vec{D}'.e^{bt}.\cos(2\pi l) + \vec{G}*(t)$$
(5)

Where $\vec{D}' = \left|\vec{G}*(t) - \vec{G}(t)\right|$ represents *i*th whales involved and the best direction, parameter b represents the constant that indicates the spiral's shape, and l represents random values from [-1, 1]. The following equation performs the WOA simulation.

$$\vec{G}(t+1) = \begin{cases} \vec{G}*(t) - \vec{A}.\vec{D} & if\ \vec{r_3} < 0.5 \\ \vec{D}'.e^{bt}.\cos(2\pi l) + \vec{G}*(t) & otherwise \end{cases}$$
(6)

In the above equation, $\vec{r_3}$ represents random values from [0,1].

The final mechanism can be achieved based on vector $\vec{A}$. Meanwhile, the random whale $\vec{G}_{rand}$ helps update the search agent's position to perform a global search for the following equation.

$$\vec{G}(t+1) = \vec{G}_{rand} - \vec{A}.\vec{D}, \vec{D} = \left|\vec{C}.\vec{G}_{rand} - \vec{G}\right|$$
(7)

Hence, the spiral process or circular movement is controlled by r_3, while the exploration and exploitation process is controlled by $\vec{A}$. The algorithm for WOA is shown step by step in Algorithm1.

Algorithm 1 Original WOA Pseudo Code

```
Initialize WOA population $\vec{G}_i(i\ 1,2,\ldots\ldots n)$ with n size, maximum iterations $Max_{iter}$ and the fitness function $F_n$.
Initialize WOA parameters $\vec{a}, \overrightarrow{A, C}, l, \vec{r}_1, \vec{r}_2, \vec{r}_3$
Initialize t as the iteration counter
Calculate fitness function $F_n$ for each $\vec{G}_i$
Find the best individual $\overrightarrow{G^*}$
while $t \le Max_{iter}$ do
    for (i=1; i < n+1) do
        if ($\vec{r_3} < 0.5$) then
            if ($|\vec{A}| < 1$), then
            Update the current search agent position using Eq.1
            else
                Select a random search agent $\vec{G}_{rand}$
                Update current search agent position by Eq.4
            end if
        else
            Update current search position by Eq. 2
        end if
    end for
    Update $\vec{a}, \overrightarrow{A, C}, l, \vec{r}_3$
    Calculate fitness function $F_n$ for each $\vec{G}_i$
    Find the best individual $\overrightarrow{G^*}$
    Set t = t+1. (increase counter).
end while
return $\overrightarrow{G^*}$
```

3.2.7 Hybrid classification with improved WOA

The proposed hybrid classification with improved WOA was performed by combining all the classification techniques, obtaining weights, and optimized by from each algorithm that has been used for the prediction, which the voting classifier has achieved. The Voting ensemble is considered a popular machine-learning approach for classification problems as it incorporates two or more learning algorithms trained with the given complete dataset. It is a machine-learning model that tries to learn from an ensemble of several independent models and predict the appropriate output class based on its highest probability [19]. There are two types of voting classifiers. One is a hard voting classifier that sees the majority votes and decides the final decision. In contrast, the other voting classifier is a soft voting classifier which decides the final decision based on the probability values of each classifier.

The procedure follows by tackling the binary classification problem holding the set of labels Y = {class 1, class 2, class 3, … class 9} and feature space X ϵ R^K where any of probabilistic classifier F holds the next function: F: X→Y. For each instance , xi, the decision profile learner j is a pair of probabilities $\left[P_{j0}, P_{j1}\right]$ that sum upto 1 while L, f, α provided parameters to the classifications. While classifying the data model with each separate classification model $C_{LR}, C_{SVM}, and\ C_{RF}$ that obtains class predictions of its own. Later the procedure follows by constructing a soft voting classifier with parameters by hyperparameter tuning and testing the model with an absolute alpha value obtained from hyperparameter tuning. The loop continues from iteration 0, the first classification model, to the last classification model, MaxIter. The Loop starts from the first classification to the last classification. For every training data point, the classification model performs training of the data, obtaining its predictive classes, which have been updated to the voting equation by counting the votes of each model. The actual class will be obtained from the majority votes, and the prediction goes to every data point in the data model.

3.2.8 Improved WOA

The improved WOA technique is the modified technique of the originally produced WOA approach. The existing method's drawbacks can be achieved in the procedure of searching; that is, instead of performing a search strategy for one random whale can be replaced and used by an advanced strategy where the whales move rapidly towards the prey, which is the best solution. From the original equation of WOA, Equation 7 intends the whales to move around themselves randomly, which is a similar mechanism of global search. While in the improved WOA, to perform enhanced exploration, a whale could follow only three whales randomly instead of one. This leads the whales to force more exploration without being affected by the leader position by replacing equation 7 with the following equation.

$$\vec{G}(t+1) = \overrightarrow{w_1} * \vec{G}_{rand\ 1} + \vec{Z} * \overrightarrow{w_2} * \left(\vec{G}_{rand\ 2} - \vec{G}_{rand\ 3}\right) + \left(1 - \vec{Z}\right) * \overrightarrow{w_3} * \left(\vec{G} - \vec{G}_{rand\ 1}\right)$$

(8)

Where $\vec{G}_{rand\ 1}\vec{G}_{rand\ 2}$, and $\vec{G}_{rand\ 3}$ are considered three random solutions, $\overrightarrow{w_1}$ is a random value that falls between [0, 0.5]. $\overrightarrow{w_2}$ and $\overrightarrow{w_3}$ are another two random values that fall between [0, 1]. $\vec{Z}$ provided to decrease exponentially instead in a linear fashion that tends to provide smooth changes between exploitation and exploration, which has been calculated using the equation,

$$\vec{Z} = 1 - \left(\frac{t}{Max_{iter}}\right)^2$$

(9)

Where t indicates the iteration number and Max_{iter} represents the maximum number of iterations it will perform.

ALGORITHM: Hybrid Classification with Improved WOA.

```
Apply SoftVot($R^k, L^0, f, \alpha$) and obtain $C_{LR}, C_{SVM}, C_{RF}$
Construct $Vote_{soft}(C_{LR}, C_{SVM}, C_{RF})$
Set iter = 0
Whileiter < MaxIter do
      For each data point
Train learner$_{x_i}$onL$_{x_i}^{iter}$
          Assign class probabilities for each $u_i \in U^{iter}$
          Perform an improved WOA function call
Get coefficients and score
          For each class
              Detect the predictive class with passing coefficients and score
              Update:
$L_{x_i}^{iter+1} \leftarrow L_{x_i}^{iter} \cup \left\{x_j, \underset{class}{\arg max} P\left(Y = class|X_{x_i}\right) \nabla\, j \in Pred_{x_i}\right\}$
$u_{x_i}^{iter+1} \leftarrow u_{x_i}^{iter} \backslash \left\{x_j\right\} \nabla_j \in Pred_{x_i}$
Iter = iter + 1
Output
Use
$Vote_{soft}(C_{LR}, C_{SVM}, C_{RF})$ with optimized weights to trained c
to predict class labels of test data.
```

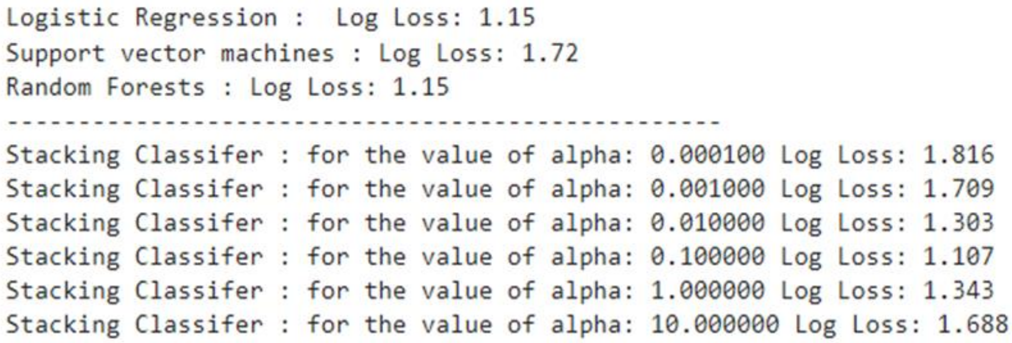
The proposed work hybrid classification with improved WOA utilized a hard voting classifier as the classification approach and improved WOA to predict the best possible results. Inside the voting classifier, the models use Logistic regression (LR), Support vector machine (SVM), and Random Forest (RF) as base classifiers in collecting votes. The model is first tuned with several hyperparameters to check the best parameter values before training the model. Later with the best alpha parameters, the model is trained by utilizing WOA-optimized weights, and the results are predicted. The results are compared with log loss metrics followed by accuracy scores with F1 measure, recall, and precision scores. The model can be utilized in real-world medical data analysis to predict the accurate cancer type before it gets into a serious issue.

4. Results and Discussions

The work utilized Google collaborator to run machine learning models in python files. The collaborator provides 12GB RAM at extreme along with 64GB storage as it gets adjusted based on the bandwidth during the execution time.

The previous work[2]provided data visualization results and complete data analysis by applying statistical methodologies like providing CDFs and PDFs graphs and performing univariate analysis on each feature to check its importance, stability, etc. The previous work [3] also performed the best algorithm checks that compared the given algorithms. It provided the Logistic Regression as the best model compared to Naïve Bayes, K- Nearest Neighbours. The work stacks the three models, Logistic

Regression, Support Vector Machine, and Random Forest, as a hybrid model approach for voting classification. Before performing the classification, the hyperparameter tuning is performed, and the results of each algorithm give specific log loss values, as shown in figure 2. Logistic regression gave a log loss value of 1.15, support vector machines obtained 1.72, while random Forest gave 1.15 as separate values. While after stacking each classifier that gave log loss as 1.107 as the least value for alpha value 0.1, thus utilizing the alpha value as 0.1, the model is trained.



```
Logistic Regression :  Log Loss: 1.15
Support vector machines : Log Loss: 1.72
Random Forests : Log Loss: 1.15
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 1.816
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 1.709
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.303
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.107
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.343
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.688
```

Fig 2. Hyperparameter Tuning

The confusion matrix for the voting classifier drawn from three main models, LR, SVM, and RF, as shown in figure 3, is found to be authentic compared to older results where classes except 8 and 9 are all bound to obtain most of the values. Similarly, in the precision and recall matrix, as shown in Figures 4 and 5, the values provided good results for class7 4, 1, and 2. For recall, too, classes 7, 6, 4, 2, and 1 show the best results, which means those classes tend to do well compared to classes 8 and 9, which have lesser values and are prone to too many misclassification points.

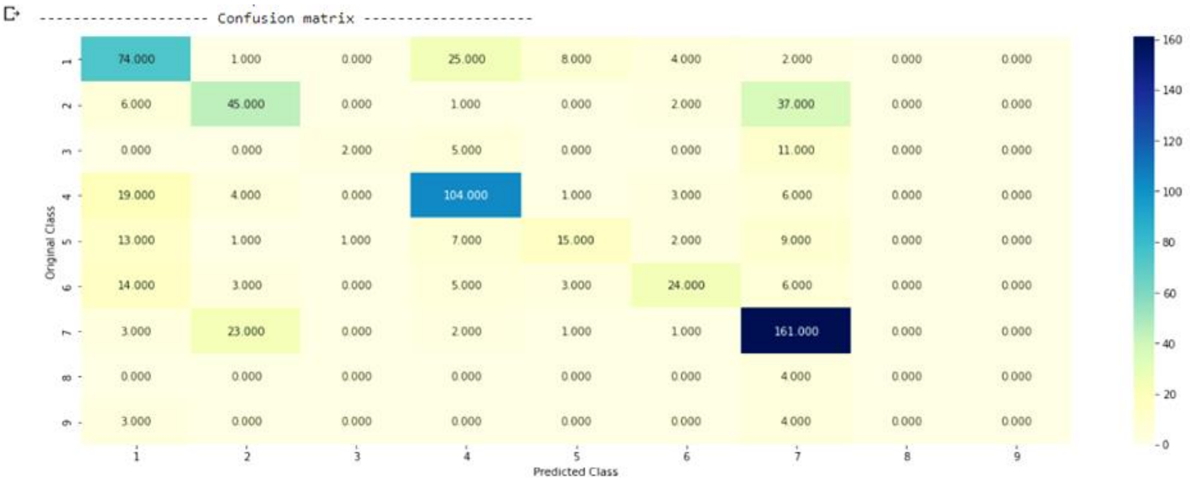


Fig 3. Confusion Matrix

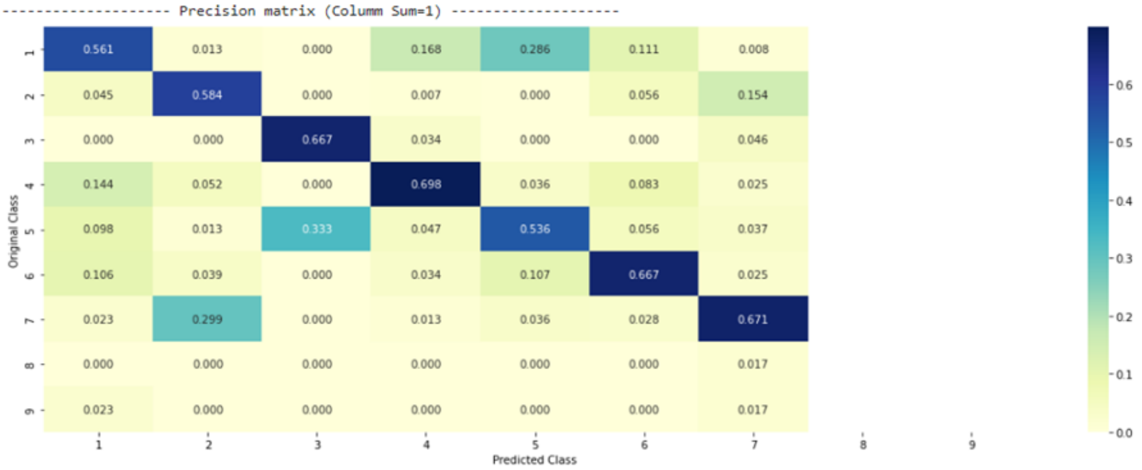
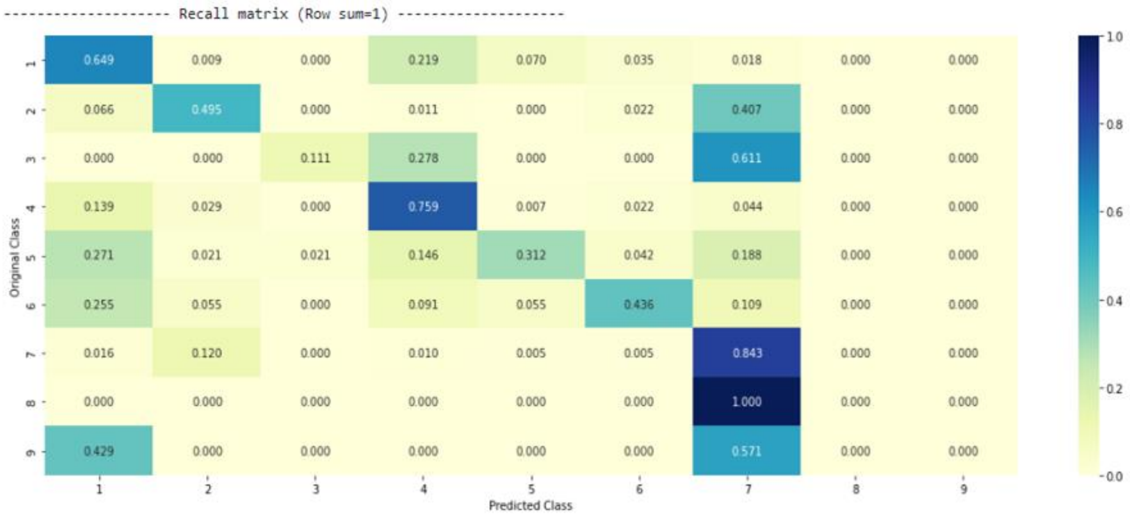


Fig 4. Precision Matrix



	1	2	3	4	5	6	7	8	9
1	0.649	0.009	0.000	0.219	0.070	0.035	0.018	0.000	0.000
2	0.066	0.495	0.000	0.011	0.000	0.022	0.407	0.000	0.000
3	0.000	0.000	0.111	0.278	0.000	0.000	0.611	0.000	0.000
4	0.139	0.029	0.000	0.759	0.007	0.022	0.044	0.000	0.000
5	0.271	0.021	0.021	0.146	0.312	0.042	0.188	0.000	0.000
6	0.255	0.055	0.000	0.091	0.055	0.436	0.109	0.000	0.000
7	0.016	0.120	0.000	0.010	0.005	0.005	0.843	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
9	0.429	0.000	0.000	0.000	0.000	0.000	0.571	0.000	0.000

Fig 5. Recall Matrix

At first, the data is well-trained with stacking individual models: Logistic Regression, Support Vector Machines, and Random Forests. While stacking up the model, each model produced individual test results showing the log loss values of their capacity. While stacking up the three models, the log loss value was 1.10. The accuracy for the stacking is also measured and found to be best as it gave 91% and precision values of 92%, while the recall value was 87% and the F-measure score was 86%. The results of the multivariate analysis of building a hybrid model were found to be best in incorporating it into the real world.

5. Conclusion

The paper provided a detailed analysis of the medical dataset specific to personalized cancer diagnosis, starting with data visualizing and comparing the machine learning models to applying hybrid models with an improved whale optimization algorithm. Among the results conducted in the previous paper as our existing system, it found that our hybrid model with improved WOA-based voting classifier stacked with Logistic Regression, Support Vector Machines, and Random Forest was found to be giving better results. In the future, the work will progress by incorporating multi-dataset techniques and approaching the same proposed hybrid model.

6. References

[1] Yan, K., & Lu, H. (2019, January). Evaluating ensemble learning impact on gene selection for automated cancer diagnosis. In International Workshop on Health Intelligence (pp. 183-186). Springer, Cham.

[2] Kandula, A. R. (2021). R, S., & S, N., Performing Uni-variate Analysis on Cancer Gene Mutation Data Using SGD Optimized Logistic Regression. International Journal of Engineering Trends and Technology, 69(2), 59-67.

[3] KANDULA, A. R., SATHYA, R., & NARAYANA, S. (2022). COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES ON GENETIC MUTATION-BASED CANCER DIAGNOSIS DATA. Journal of Theoretical and Applied Information Technology, 100(6).

[4] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology, 110, 12-22.

[5] Wei, L., Wan, S., Guo, J., & Wong, K. K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artificial intelligence in medicine, 83, 82-90.

[6] Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an efficient intrusion detection system based on feature selection and ensemble classifier. Computer networks, 174, 107247.

[7] Pratama, M., Pedrycz, W., & Lughofer, E. (2018). Evolving ensemble fuzzy classifier. IEEE Transactions on Fuzzy Systems, 26(5), 2552-2567.

[8] Qiu, W. R., Xiao, X., Xu, Z. C., & Chou, K. C. (2016). iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget, 7(32), 51270.

[9] Yousaf, A., Umer, M., Sadiq, S., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2020). Emotion recognition by textual tweets classification using the voting classifier (LR-SGD). IEEE Access, 9, 6286-6295.

[10] Atallah, R., & Al-Mousa, A. (2019, October). Heart disease detection using machine learning majority voting ensemble method. In 2019 2nd international conference on new trends in computing sciences (ictcs) (pp. 1-6). IEEE.

[11] Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. Journal of Imaging, 6(6), 39.

[12] El-Kenawy, E. S. M., Ibrahim, A., Mirjalili, S., Eid, M. M., & Hussein, S. E. (2020). Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. IEEE Access, 8, 179317-179335.

[13] Sreedhar P, Siva Satya & Kandula, Ashok & Tamilarasi, K & Maan, Srikrishna & Alsekait, Deema & Publishing House, International Scitech. (2022). Deep Neural

Network for Image Recognition In Medical Diagnosis. Journal of Pharmaceutical Negative Results. 2022. 10.47750/pnr.2022.13.S09.47.

[14] Goyal, A., & Sardana, N. (2019, August). Empirical analysis of ensemble machine learning techniques for bug triaging. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.

[15] Jan, Z. M., & Verma, B. (2019, June). Ensemble classifier optimization by reducing input features and base classifiers. In 2019 IEEE Congress on Evolutionary Computation (CEC) (pp. 1580-1587). IEEE.

[16] Dogan, A., & Birant, D. (2019, September). A weighted majority voting ensemble approach for classification. In 2019 4th International Conference on Computer Science and Engineering (UBMK) (pp. 1-6). IEEE.

[17] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer genomics & proteomics, 15(1), 41-51

[18] Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 6308-6325

[19] Dogan, A., & Birant, D. (2019, September). A weighted majority voting ensemble approach for classification. In 2019 4th International Conference on Computer Science and Engineering (UBMK) (pp. 1-6). IEEE.

[20] Swarajya lakshmi v papineni, A.Mallikarjuna Reddy, Sudeepti yarlagadda , Snigdha Yarlagadda, Haritha Akkineni "An Extensive Analytical Approach on Human Resources using Random Forest Algorithm" International Journal of Engineering Trends and Technology 69.5(2021):119-127.

[21] A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha, "Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 4-Regular Issue, 2018.

[22] Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G., et al., (2019), "Generating cancelable fingerprint template using triangular structures", Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, pp. 1951-1955(5), doi: https://doi.org/10.1166/jctn.2019.7830.

[23] Mallikarjuna Reddy, A.,Venkata Krishna, V. and Sumalatha, L." Face recognition approaches: A survey" International Journal of Engineering and Technology (UAE), 4.6 Special Issue 6, volume number 7 , 117-121,2018.

[24] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, "Leveraging bio-maximum inverse rank method for iris and palm recognition", International Journal of Biometrics, 2022 Vol.14 No.3/4, pp.421 - 438, DOI: 10.1504/IJBM.2022.10048978.

[25] V. NavyaSree, Y. Surarchitha, A. M. Reddy, B. Devi Sree, A. Anuhya and H. Jabeen, "Predicting the Risk Factor of Kidney Disease using Meta Classifiers," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972392.

[26] B. H. Rao et al., "MTESSERACT: An Application for Form Recognition in Courier Services," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), 2022, pp. 848-853, doi: 10.1109/ICOSEC54921.2022.9952031.

[27] P. S. Silpa et al., "Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm," 2022 3rd International Conference *on Smart Electronics and* Communication (ICOSEC), 2022, pp. 759-767, doi: 10.1109/ICOSEC54921.2022.9951883.

[28] B. S. Reddy, A. Mallikarjuna Reddy, M. H. D. S. Sradda, T. Mounika, S. Mounika and M. K, "A Comparative Study on Object Detection Using Retinanet," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972742.