International Journal of Applied and Scientific Research (IJASR)
Vol. 2, No. 2, 2024 : 263-286

Application of Stacked Ensemble Techniques on Ensemble Feature Selection Techniques for Classifying Recurrent Head and Neck Squamous Cell Carcinoma Prognosis

Joseph Acquah[1]*, Damianus Kofi Owusu[2]
University of Mines and Technology, Tarkwa, Ghana
Corresponding Author: Joseph Acquah jacquah@umat.edu.gh

A R T I C L E I N F O

Keywords: Recurrent Head and Neck Cancer, Ensemble Feature Selection, Stacking, Classification

Received : 15, December
Revised : 18, January
Accepted: 23, February

A B S T R A C T

This study aimed to identify the optimal combination of the stacked ensemble (SE) and the heterogeneous ensemble feature selection (HETR-EFS) technique for classifying HNSCC recurrence patterns. Four SE classification models were developed based on various EFS techniques, using GBM meta-classifiers in each case. The results showed that implementing the SE technique consisting of five base classifiers on the heterogeneous ensemble feature (HETR-EF) subset achieved better performance than other EF subsets and HETR-EFs. Thus, learning SE technique having five base classifiers on HETR-EFs is clinically appropriate as a prognostic model for classifying and predicting HNSCC patients' recurrence data. The SE technique, which combines base classifier models, is clinically appropriate for classifying and predicting HNSCC patients' recurrence data. The study highlights the importance of finding a machine learning algorithm that performs best given varied distributions, as not all algorithms are equally created.

INTRODUCTION

For successful and complete destruction of malignant cells in the body, the treatment of recurrent head and neck squamous cell carcinoma (HNSCC) requires correct prognosis linked with it in order to define the kind and extent of the therapy. In the interim, many prognostic models based on clinical and histopathologic parameters for recurrent HNSCC have been researched and developed, not from a medical perspective but from a scientific point of view in different fields using statistical methods, Artificial Intelligence (AI), and ML techniques; addressing the issue of the patient's disease recurrence [5,9]. In medicine, a patient's disease is determined by its signs and symptoms (called a diagnosis) and the prognosis is the study of how the disease will affect the patient. Cancer has been classified as a heterogeneous disease with various subgroups. The goal of applying ML approaches has been to construct a model for the progression and management of cancer subtypes. Various machine learning (ML) techniques, such as but not limited to Artificial Neural Networks (ANNs), Random Forest (RF), GBM, NB, and GLM, have been applied in a wide range of cancer research to build predictive models from complex datasets. These models are known to offer effective and high accuracy in decision making, highlighting their importance. Even though the majority of these ML approaches produce some useful results, they fall short of the high accuracy requirements for predicting the complex cancer environment, and they also exhibit insufficient generalization ability when predicting the labels of upcoming unobserved data or scenarios. Finding a classifier model that successfully predicts and classifies the labels of future, ambiguous data is the aim of classification. Therefore, a classification model should not overfit the training data; rather, it should be general enough to encompass previously undiscovered situations.

However, ensemble learning, which turns a number of base classifier models into a strong one by combining them, may produce a strong classifier with good generalization ability. It is extremely difficult to get a single classification model for both feature selection and model training with this ability. But in contrast to ensembles, bagging and boosting, stacking, or stacked generalization is the most successful ensemble learning technique, according to], if ensemble learning can improve the classification model's generalization capacity. This method of heterogeneous ensemble learning uses a meta-learning algorithm to aggregate the strongest set of numerous base learners into a strong learner. In fact, hold the opinion that this method (stacking ensemble learning) has been observed to yield more accurate findings in numerous techniques and studies for which it has been used. The integration of stacked ensemble technique with heterogeneous ensemble feature selection technique is the only option to fully overcome these drawbacks and build a classification model with strong generalization ability.

One of the key issues in machine learning is feature selection (FS). Analysis of a classifier's ability to predict a label and the features it makes use of can help researchers learn more about numerous applications, such as bio-informatics or neurology. Additionally, efficient feature selection creates

classifiers that are parsimonious, need less memory, and train and test more quickly. It can also lower feature extraction costs and improve generalization ability. When looking for linear dependencies between features and labels, linear feature selection algorithms like LARS are quite effective. When characteristics interact nonlinearly, they fall short though. Nonlinear interactions can be handled by nonlinear homogeneous ensemble feature selection techniques like random forest [or recently developed kernel methods [30,24]. But when the size of the training set increases, its computational and memory cost often increases super-linearly. This becomes more of an issue as datasets get bigger. Scalability and nonlinear feature selection must be balanced; however, this is still an unsolved issue. A feature selection method should, in theory, be able to extract pertinent features with high reliability, recognize non-linear feature interactions, scale linearly with the number of features and dimensions, and take into account known sparsity structures. This study investigated how the stacked ensemble techniques of varying base classifiers can be employed in the prognosis of HNSCC recurrence based on ensemble (heterogeneous or homogenous) features, provided by some selected EFS techniques by building a prognostic model able to classify the recurrent for HNSCC prognosis. The evaluation of these stacked ensemble models on various ensemble feature subsets of varying difficulty and size, and the demonstration that HETR-EFS tends to match or outperform the accuracy and feature selection trade-off of random forest FS, the current state-of-the-art in nonlinear feature selection-gradient boosted feature selection (GBM-EFS). The study showedcase the ability of HETR-EFS to naturally incorporate side-information about inter-feature dependencies on a real-world biological classification task.

LITERATURE REVIEW

Several studies in the domain of cancer cases have been studied using stacking ensemble learning techniques. Some of which are: proposed a one-layer stacked ensemble-based model (with 10-fold CV) having two single base classifiers; NB and SVM, for the classification of recurrent breast cancer prognosis where DT was used as a meta-classifier to stack base classifiers. developed a one-layer stacking ensemble technique having three single base classifiers; KNN, NB, and DT (C4.5), and GLM as a meta-learner, able to predict the types of cancer around the HNC regions (Sinonasal, nasopharyngeal, laryngeal, and thyroid) applied the same technique (KNN, NB, and DT (C4.5) as base learners, and GLM as a meta-learner) in the diagnosis of HNC susceptibility to facilitate prompt referral.

To generate stacked ensemble model proposed a stacking ensemble-based algorithm, a technique that found the optimal weighted average of diverse base learners for classification of various healthcare datasets (Wisconsin Breast Cancer, Pima Indian Diabetes Dataset, and Indian Liver Patient Dataset using GBM, DRF, and DNN as base learners, and GLM as a meta learner. Their techniques consisted of stacking or super learner having two base classifiers (GBM and DRF) and that having three base classifiers (GBM, DRF, and DNN)

and concluded that a super learner having three base learners outperformed that having two base learners. Thus, their recommendation was that, future study should investigate by including diverse base learners and meta-learners in stacking ensemble for various healthcare datasets. Based on this recommendation, [19] proposed a stacking ensemble-based algorithm, a technique that found the best meta-learner in a stacking ensemble for classifying breast cancer, using GBM, DRF, DNN, and GLM as base learners in a stacking ensemble, each of which was re-learned as a meta-learner to determine the best meta-learner in the stacking ensemble having four base learners. Their study showed that using specific models as a meta-learner resulted in better performance than single classifiers, and using GBM or GLM as a meta-learner is appropriate as a supporting tool for classifying breast cancer data. Thus, the overall purpose of the present study was to develop a stacking ensemble classification model as a supporting tool that combines weak/base ensemble classifiers and single base classifiers needed for robust prognosis for early diagnosis and treatment outcomes based on the optimal feature subset of clinical, histopathologic (pathologic) and genomic markers, including other risk factors and treatment types associated with HNSCC recurrence in Ghana for accurate prognosis. There has not been any study yet on recurrent HNSCC prognosis using the same technique or an adapted stacking ensemble technique in Ghana. Base on the ML algorithms considered by [19] as the most effective algorithms to providing the most effective ensemble classification model for HNSCC prognosis, all have been employed under this study with the inclusion of NB to experiment a stacked ensemble consisting of five (5), at least one more than that of the state-of-the-art stacked ensemble model consisting of a maximum of four (4) base classifiers in HNC prognosis. Thus, NB was chosen from among the most effective single base classifiers (DT, KNN, NB, and SVM) considered by the previous studies, based on its performance on the experimental data to experiment more than a maximum of five base classifiers in a stacking ensemble.

Several studies in the domain of cancer cases have been studied using stacking ensemble learning techniques. Some of which are: [2] proposed a one-layer stacked ensemble-based model (with 10-fold CV) having two single base classifiers; NB and SVM, for the classification of recurrent breast cancer prognosis where DT was used as a meta-classifier to stack base classifiers. [4] developed a one-layer stacking ensemble technique having three single base classifiers; KNN, NB, and DT (C4.5), and GLM as a meta-learner, able to predict the types of cancer around the HNC regions (Sinonasal, nasopharyngeal, laryngeal, and thyroid). [3] applied the same technique (KNN, NB, and DT (C4.5) as base learners, and GLM as a meta-learner) in the diagnosis of HNC susceptibility to facilitate prompt referral.

To generate stacked ensemble model, [18] proposed a stacking ensemble-based algorithm, a technique that found the optimal weighted average of diverse base learners for classification of various healthcare datasets (Wisconsin Breast Cancer, Pima Indian Diabetes Dataset, and Indian Liver Patient Dataset using GBM, DRF, and DNN as base learners, and GLM as a meta learner. Their

techniques consisted of stacking or super learner having two base classifiers (GBM and DRF) and that having three base classifiers (GBM, DRF, and DNN) and concluded that a super learner having three base learners outperformed that having two base learners. Thus, their recommendation was that, future study should investigate by including diverse base learners and meta-learners in stacking ensemble for various healthcare datasets. Based on this recommendation, [19] proposed a stacking ensemble-based algorithm, a technique that found the best meta-learner in a stacking ensemble for classifying breast cancer, using GBM, DRF, DNN, and GLM as base learners in a stacking ensemble, each of which was re-learned as a meta-learner to determine the best meta-learner in the stacking ensemble having four base learners. Their study showed that using specific models as a meta-learner resulted in better performance than single classifiers, and using GBM or GLM as a meta-learner is appropriate as a supporting tool for classifying breast cancer data. Thus, the overall purpose of the present study was to develop a stacking ensemble classification model as a supporting tool that combines weak/base ensemble classifiers and single base classifiers needed for robust prognosis for early diagnosis and treatment outcomes based on the optimal feature subset of clinical, histopathologic (pathologic) and genomic markers, including other risk factors and treatment types associated with HNSCC recurrence in Ghana for accurate prognosis. There has not been any study yet on recurrent HNSCC prognosis using the same technique or an adapted stacking ensemble technique in Ghana. Base on the ML algorithms considered by [19] as the most effective algorithms to providing the most effective ensemble classification model for HNSCC prognosis, all have been employed under this study with the inclusion of NB to experiment a stacked ensemble consisting of five (5), at least one more than that of the state-of-the-art stacked ensemble model consisting of a maximum of four (4) base classifiers in HNC prognosis. Thus, NB was chosen from among the most effective single base classifiers (DT, KNN, NB, and SVM) considered by the previous studies, based on its performance on the experimental data to experiment more than a maximum of five base classifiers in a stacking ensemble.

Bagging

Bagging, sometimes referred to as Bootstrap Aggregation, is an ensemble machine learning technique that combines weaker base learners into a stronger learner. Bootstrapping is the process of creating replacement datasets at random and using these varied random subsets of the data to train various classifiers. Bagging is the term used to describe the process of using this technique to integrate different classifiers (decision trees). As a result, bagging simply refers to building each classifier or tree using a unique random subset of the dataset that is drawn via replacement. To create the final forecast, the predictions from each independent classifier might be averaged (regression) or decided upon by a majority (classification). Random Forest (RF) is an algorithm or method that is frequently utilized. A model that overfits the training data will have its complexity reduced using the ensemble-based technique random

forest [8, 23]. Random forest is explored and employed in the study's feature selection and classification model learning processes.

Table 1. Bagging Algorithm

	Algorithm 1: Bagging Algorithm
Input	Training set $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ $(\mathbf{x}_i \in \mathbb{R}^m, y_i \in Y)$
Output	An ensemble classifier H
1:	**for** $t \leftarrow 1$ to T **do**
2:	Construct a sample data D_t by randomly sampling with replacement in D
3:	Learn a base classifier h_t based on D_t
4:	**end for**
5:	**return** $\boldsymbol{H}(\mathbf{x}) = \underset{y \in Y}{\operatorname{argmax}} \sum_{t=1}^{T} \mathbf{1}(h_t(\mathbf{x}) = y)$

Boosting

In order to reduce training error, boosting is an ensemble learning method for homogeneous learning that combines a homogenous group of weak learners into a strong learner. In boosting, a randomly chosen sample of data is chosen, fitted with the learner, and then consecutively learned. In other words, each student seeks to make up for the shortcomings of its elder. One strong prediction rule is created by combining the weak rules from each learner during each cycle. The three widely used approaches of adaptive, gradient, and extreme gradient boosting (AdaBoost, GradientBoost, and XGBoost) are the main emphasis of the strategies for boosting [29]. GradientBoost, or GBM, is discussed and used for feature selection and learning a classification model for the study's purposes.

Table 2. Boosting Algorithm

	Algorithm 2: Boosting Algorithm
Input	Training set $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ $(\mathbf{x}_i \in \mathbb{R}^n, y_i \in \{+1, -1\})$
Output	An ensemble classifier H
1:	Initialize the weight distribution W_1
2:	**for** $t \leftarrow 1$ to T **do**
3:	Learn weak classifier h_t based on D and W_t
4:	Evaluate weak classifier $\varepsilon(h_t)$
5:	Update weight distribution W_{t+1} based on $\varepsilon(h_t)$
6:	**end for**
7:	**return** $\boldsymbol{H} = combination(\{h_1, \ldots, h_T\})$

Stacking

Stacking is a technique that combines heterogeneous of multiple weak/base learners into a more robust learner than individual base learners. This technique combines the predictions of different individual base learners to make a final robust prediction. Where weak or base learning algorithms are rightfully blended, a meta-learner with lower variance and bias can be

developed. Stacking uses cross-validation to estimate the performance of multiple base learning algorithms. The output from the base learners, called "level-one" data in the stacking literature, serves as input to the meta-learning algorithm(s). Stacking learns a high-level classifier on top of the base classifiers. It can be viewed as a form of meta learning where the base classifiers, also known as first-level classifiers, are combined to train a second-level classifier, also known as a meta-classifier. Based on the literature review and the study's objectives, the base classifiers GBM, DRF, DNN, NB, and GLM were employed for feature selection and classification learning.

Table 3. Stacking

	Algorithm 3: Stacking
Input	Training set $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ $(\mathbf{x}_i \in \mathbb{R}^m, y_i \in Y)$
Output	An ensemble classifier H
1:	Step 1: Learn first-level classifiers
2:	**for** $t \leftarrow 1$ to T **do**
3:	Learn a base classifier h_t based on D
4:	**end for**
5:	Step 2: Construct new data sets from D
6:	**for** $i \leftarrow 1$ to n **do**
7:	Construct a new data set that contains $\{\boldsymbol{x}_i', \mathrm{y}_i\}$, where $\mathrm{x}_i' = \{h_1(\mathbf{x}_i), h_2(\mathbf{x}_i), \dots, h_T(\mathbf{x}_i)\}$.
8:	**end for**
9:	Step 3: Learn a second-level classifier
10:	Learn a new classifier h' based on the newly constructed data set
11:	**return** $\boldsymbol{H}(\mathbf{x}) = h'\big(h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x})\big)$

Proposed Approach: Multi-Level Stacked Ensemble Learning

For the purpose of this study, a novel approach for classifying and predicting recurrent HNSCC prognosis was proposed in HNCs environment. This approach extended the existing stacking techniques of as discussed in the literature by improving in the areas of:

- ✓ Strong and more diverse base classifiers against small number of classifiers as used in previous studies.
- ✓ More diverse meta-classifiers against small number of meta-classifiers classifiers as used in previous studies.
- ✓ Combining meta-classifier models with heterogeneous ensemble feature selectors against the previous existing approach where no such combination was not learned.

The detail description of the technique is explained as follows: Using GBM, DRF, DNN, NB, and GLM techniques for feature selection, by ordering the features according to their importance, in order to provide optimal feature subset, consider a labeled dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ on n patients (instances) with feature vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Consider also an ensemble

EFS = $\{FS_1, FS_2, \ldots, FS_t\}$ which consists of t base feature selectors (FS), where t is the number of feature selectors. Each feature selector FS_i provides a feature subset $FS_i = \{x_1^i, x_2^i, \ldots, x_{n-1}^i\}$, where n-1 is the number of selected features by the i^{th} feature selection method. To implement the aggregation in EFS technique, the sum of the subsets generated by t FS algorithms is estimated according to equation (1) or (2). For each feature j in the subset SUM, compute an index (weight) of importance according to equation (3), and obtain the weighted feature subset SUMW according equation (4), where m is number of features in summation. The feature j importance is determined by the ratio of the number of times it is present in the feature subset SUM to t FS algorithms. Sort the m features according to their importance. Finally, based on the threshold α, select the α% features (features that exceed a threshold) from the feature ordered according to their importance to obtain the optimal feature subset $\mathrm{SUM}_{\mathrm{best}} = \mathbf{x}' = (x_1', x_2', \ldots, x_{n-1}')$, so that, the new dataset becomes $D_{new} = \{\mathbf{x}_i', y_i'\}_{i=1}^n$.

$$\mathrm{SUM} = \sum_{i=1}^{t} FS_i \qquad (1),$$

$$\mathrm{SUM} = \sum_{m=1}^{t} k_1 x_1^t + k_2 x_2^t + \cdots + k_m x_m^t \qquad (2)$$

$$w_j = \frac{k_j x_j^t}{t} \qquad (3),$$

$$\mathrm{SUM}W = \{w_1 x_1^t + w_2 x_2^t + \cdots + w_m x_m^t\} \qquad (4)$$

Where w_j is the weight of the feature j, and k is the number of times the feature j is present in subset SUM.

This paper presented four different techniques of stacked ensemble learning on two different techniques of ensemble feature selection: one being the heterogeneous EFS (the combination of GBM, DRF, DNN, GLM, and NB) and the other one being homogeneous EFS. The first stacked ensemble learning used two base classifiers, namely gradient boosting machine (GBM) and distributed random forest (DRF); the second one used three base classifiers, namely GBM, DRF, and deep neural network (DNN); the third one used four base classifiers, namely GBM, DRF, DNN, and generalized linear model (GLM); and the fourth one used five base classifiers, namely GBM, DRF, DNN, GLM, and Naïve bayes (NB); and in each case, a meta-classifier called GBM was used. Various cancer data subsets related to HNSCC provided by various feature selection techniques used in this study were used, and compare the performance of stacked ensemble models on these various data subsets. The evaluation results confirmed that stacked ensemble techniques built on Gradient Boosted feature subset (GBM-FS) has the ability to perform better compared to stacked ensemble techniques built on feature subsets provided by other feature selection techniques. Similarly, the evaluation results confirmed that stacked ensemble techniques consisting of five base classifiers has the ability to perform better compared to other stacked ensemble techniques considered on five feature subsets of HNSCC dataset. To achieve better

performance using these base classifiers from H2o, GBM, DRF, DNN, GLM, and NB were selected. For the meta-classifier, GBM model was used as it was the best performing base classifier among the base classifiers considered in this study as shown in Figure 1. To obtain data subsets for learning stacked ensemble techniques, each base classifier was used to perform feature selection, each of which ranked the features according to their importance; and using 80% threshold, feature subsets were obtained as shown in Table 4 and Table 5 respectively. The Algorithm 4 shows the learning of stacked ensemble models with 10-fold cross-validation.

Table 4. Stacking with K-fold (K=10) Cross Validation

Algorithm 4 Stacking with K-fold (K=10) cross validation	
Input:	Dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$; learning rate $\alpha > 0$
	Ensemble feature subset $D_{new} = \{\mathbf{x}_i', y_i'\}_{i=1}^{n}$
	D_s - Training set
	D_t - Testing set
	EFS $= \{FS_1, FS_2, \dots, FS_t\}$– feature selection algorithms which constitute the ensemble feature selection.
	$C = \{h_1, h_2, \dots, h_L\}$ – classifiers set which constitute the ensemble.
Output:	An ensemble classifier $\boldsymbol{H}$
/*Phase I: Ensemble Feature Selection*/	
Step 1: Get feature sets by different feature selection algorithms	
for algorithm FS_i in $\{FS_1, FS_2, \dots, FS_t\}$	
Using dataset D do feature selection by algorithm FS_i	
for $i \leftarrow 1$ to t **do**	
Sum the subsets of t feature selection algorithms	
end for	
Step 2: Get weight sequence of t feature selection algorithms	
for FS_i in $\{FS_1, FS_2, \dots, FS_t\}$	
for x_j in SUM= $\{FS_1 + FS_2 + \cdots + FS_t\}$	
$\omega_j = k_j x_j^t / t$	
Return the feature weight sequence W	
Step 3: Get best feature subsequence according to α	
Sorted FS according to W	
Put the first α% features in SUM to SUM_{best} ($\text{SUM} \rightarrow \text{SUM}_{\text{best}}$)	
Return SUM_{best}	
/*Phase II: Training*/	
Step 4: Adopt cross validation approach in preparing a training set for meta-classifier	
Randomly split D_s into V equal-size subsets: $D = \{D_1, D_2, \dots, D_K\}$	
for $v \leftarrow 1$ to K **do**	
Step 4.1: Learn first-level classifiers $\{h_1, h_2, \dots, h_L\}$	
for $l \leftarrow 1$ to L **do**	
Learn a classifier h_{kl} from D/D_k	
end for	
Step 4.2: Construct a training set for second-level classifiers	
for $\mathbf{x}_i' \in D_k$ **do**	
Get a record $\{\mathbf{x}_i'', \mathbf{y}_i'\}$, where $\mathbf{x}_i'' = \{h_{k1}(\mathbf{x}_i'), h_{k2}(\mathbf{x}_i'), \dots, h_{kL}(\mathbf{x}_i')\}$	
end for	
end for	
Step 5: Learn second-level classifier	

Re-learn first-level classifier h'_l from the collection of $Z = \{\mathbf{x}''_i, y'_i\}_{i=1}^{n}$
end for
Return $H(\mathbf{x}) = h'\left(h_1(\mathbf{x}'), h_2(\mathbf{x}'), \ldots, h_L(\mathbf{x}')\right)$
/*Phase III: Evaluation*/
Step 6: Predict unseen example (testing set)
for each $\mathbf{x} \in D_t$ **do**
Apply an ensemble classifier $H(\mathbf{x}')$ on $\mathbf{x}'$.
end for

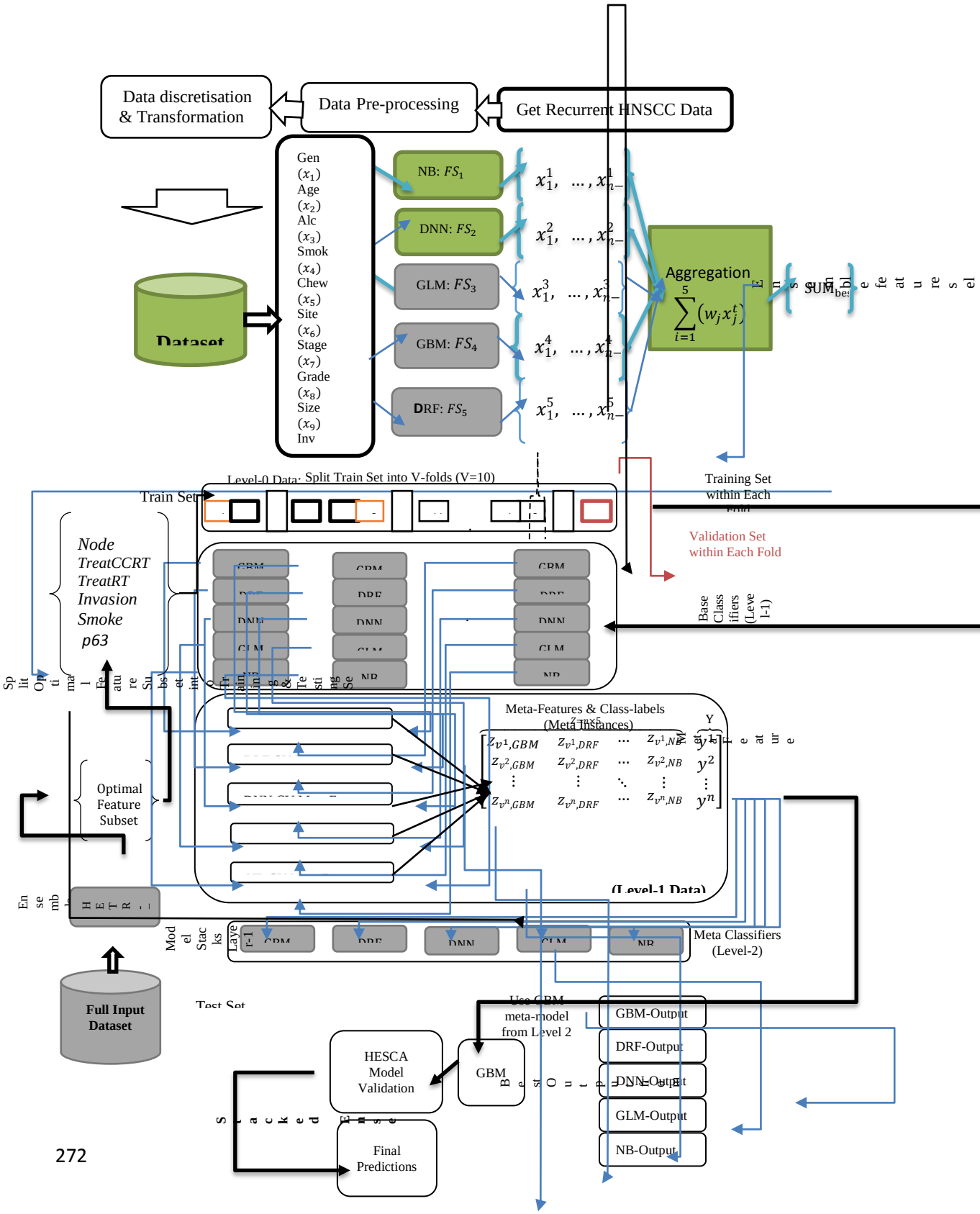


Figure 1: Architecture of Stacked Ensemble Model

METHODOLOGY

This research sought to discover the most effective pairing of stacked ensemble (SE) and heterogeneous ensemble feature selection (HETR-EFS) methods in classifying recurrence patterns in Head and Neck Squamous Cell Carcinoma (HNSCC). Four classification models were created by combining SE with different EFS techniques, incorporating GBM meta-classifiers in each instance.

RESEARCH RESULT AND DISCUSSION

Dataset

To evaluate the performance of the stacked ensemble classification models, a retrospective cohort study of 125 HNSCC patients out of the population of 185 patients aged ≥15 years, previously diagnosed of HNSCC subtypes including laryngeal cancer, hypopharyngeal cancer, nasopharyngeal cancer, and oropharyngeal cancer and treated with curative intent at KBTH, where cancer reached remission but overtime, had either recurrence or nonrecurrence between 2016 and 2020 were sampled. For each patient, information on his/her Gender, Age at diagnosis, Alcohol drinking habit, Smoking habit, Quid chewing habit, Primary site of tumor, Tumor stage at diagnosis, Histological grade, Tumor size, Depth of invasion front, Cervical lymph/Neck nodes, Pathological tumor staging, Pathological lymph nodes, Family history of cancer, Human papillomavirus level, *p16* type, *p63* type, and type of treatment are taken into consideration. The dataset has a total of 125 instances, 18 attributes (features), and a class label with binary outcome coded 1 (as recurrence) or 0 (as nonrecurrence). There are 33 and 92 female and male records respectively. The summary of this dataset is shown Table 1. In medical research, it takes time to collect sufficient samples as most patients are usually lost to follow-up to check whether or not they had a recurrence and so, the sample size is usually small. HNSCC is considered recurrence if the patient was treated with curative intent and after the cancer reaches its remission, they redeveloped HNSCC termed as recurrence. Patients that received palliative treatment intent and still had cancer are not considered cancer recurrent patients. Unfortunately, most patients received palliative intent treatment and only a few could receive curative intent due to financial difficulties, causing small number of instances. The number of features in the dataset is considered too many (18 attributes) if compared to the sample size (125 instances). Thus, the feature selection method is needed to reduce the number of features and select only those that are significant to the classification model. The original dataset was subjected to five feature selection techniques, namely GBM, DRF, DNN, GLM, and NB, each provided feature subset of the data as shown Table 4. Training data (75%) and test data (25%) were constructed for each data subset. A machine learning library for R programming language was used. Data augmentation was generally used to improve a model's performance. This is a

technique that comprises a set of methods used to artificially increase the number of data samples present in the dataset. This was done as deep learning models generalize well when the number of data samples available to train on is large. In this way, state-of-the-art models can be created with fewer data samples available. The data augmentation technique is usually applied to computer vision applications where domain-specific data, such as medical data, is not abundantly available. Thus, the usage of data augmentation technique.

Table 5. Dataset description

Dataset	No. of instances	No. of attributes	Class label with No. of instances
HNSCC	125	18	Class 1: recurrence (61); class 0: nonrecurrence (60)

Data Pre-Processing

A normalised predictive mode approach was used to identify and fill the missing instances. This approach of imputation is suitable for categorical (nominal) data; therefore, this technique is feasible in this study as the size of training examples is very small without needing to discard or delete the case having missing training instances under any feature. For data discretisation and transformation, one hot-encoding was used for features with more than two levels in order to have a normalised dataset for training, evaluation and prediction. So that, the initial 18 number of features then became 35 number of features in the dataset to be considered for learning. Table 1 presents a description of HNSCC dataset and Table 6 presents feature subsets that were ready for ingestion into the process of model training and evaluation.

Performance Metrics

To measure the performance of a classification model on the ensemble feature subsets of recurrent HNSCC prognosis dataset, the most commonly used performance measures of accuracy in cancer prognosis were utilised. These performance metrics are; accuracy, logloss, recall, specificity, and Area Under Receiver Operating Characteristic Curve (AUROC).

Table 6. Confusion Matrix for Recurrent HNSCC Prognosis

		Actual conditions	
		Recurrence (Positive)	Nonrecurrence (Negative)
Predicted outcomes	Recurrence (Positive)	True positive (TP)	False positive (FP)
	Nonrecurrence (Negative)	False negative (FN)	True negative (TN)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \times 100\% \quad (5)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \times 100\% \quad (6)$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (7)$$

Unlike AUC which looks at how well a model can classify a binary target; Logarithmic loss (log loss) metric evaluates how close a model's predicted values are to the actual/true value (0 or 1 in case of binary classification). That is, it measures the uncertainty of the predicted labels based on how far it varies from the actual label. The more the predicted probability diverges from the actual value, the higher is the log loss value. Thus, a lower log loss value means better predictions.

Log loss equation for binary classification is;

$$\text{Logloss} = -\frac{1}{N}\sum_{i=1}^{N} w_i \left(y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)\right) \quad (8)$$

Where:

- N is the total number of rows (observations) of the corresponding
- N is the total number of rows (observations) of the corresponding dataframe
- w is the per row user-defined weight (default is 1)
- p is the predicted value assigned to a give row (observation)
- y is the actual target value

Many classifiers can be made to work off a threshold. For example, in cancer diagnosis, if the diagnostic scores computed by the predictive model for certain observations cross a certain threshold, then these observations may be deemed to be cancerous. In the context of such threshold-based classifiers, a single precision/recall metric may provide the complete performance profile. The precision and recall are computed for a particular assignment of examples into positive and negative classes. By changing the threshold, the assignment changes. The Receiver Operating Characteristic (ROC) curve allows the evaluation of such a complete performance profile of the model. The ROC curve is the plot of true positive rate (sensitivity/recall) versus the true negative rate (specificity) for changing values of the threshold. The ROC curve is comprehensive. It can be condensed to a single value termed as the AUC. AUC is simply the area calculated under the ROC curve; a single value and lies within 0 and 1. The larger the area under the curve the better is the prediction (Adbul-Kareem, 2002). So, higher the value of AUC, better is a classifier and vice versa.

Table 7. Classifiers with their corresponding hyper-parameter values

Classifiers	Hyper-parameters in grid search with the corresponding range of values	Hyperparameters fixed values
GBM	max_depth = c(7, 9), learn_rate = c(0.01, 0.1), learn_rate_annealing=c(0.99, 1), sample_rate=c(0.5, 0.7, 1), col_sample_rate=c(0.8, 0.9, 1)	ntrees = 3000 nfolds = 10 fold_assignment = "Modulo" keep_cross_validation_predictions = TRUE stopping_rounds = 50
DRF	max_depth = c(9, 30), mtries = 3, sample_rate = c(0.5, 0.75, 1), col_sample_rate_per_tree= (0.8, 0.9, 1)	ntrees = 3000 nfolds = 10 fold_assignment = "Modulo" keep_cross_validation_predictions = TRUE stopping_rounds = 50
DNN	activation=c("Rectifier", "Tanh"), hidden = c(5, 10, 50), l1 = c(0, 1e-3, 1e-5), l2 = c(0, 1e-3, 1e-5),	epochs = 20 nfolds = 10 fold_assignment = "Modulo" keep_cross_validation_predictions = TRUE stopping_rounds = 50
NB	laplace=c(0, 5, by 0.5)	nfolds = 10 fold_assignment = "Modulo" keep_cross_validation_predictions = TRUE
GLM	alpha=c(0.1)	nfolds = 10 remove_collinear_columns = TRUE fold_assignment = "Modulo" keep_cross_validation_predictions = TRUE

DISCUSSION

This study compared the performance of stacked ensemble techniques implemented on various feature subsets of the HNSCC dataset provided by various ensemble feature selection techniques used in this study. The stacked ensemble techniques were trained on the training set, and were evaluated on the test set for each data subset. Table 7 shows the performance of the proposed stacked ensemble technique having two base classifiers (GBM and DRF) on the test set for different feature subsets of the data; Table 8 shows the performance of the proposed stacked ensemble technique having three base classifiers (GBM, DRF, and DNN) on the test set for all the data subsets used in this study; Table 9 shows the performance of the proposed stacked ensemble technique having four base classifiers (GBM, DRF, DNN, and GLM) on the test set for all the data subsets used in this study. while Table 10 shows the performance of the proposed stacked ensemble technique having five base classifiers (GBM, DRF, DNN, GLM, and NB) on the test set for all the data subsets used in this study.

Table 8. Top 20 Most Important Features (out of 35) by Ensemble Feature Selection

	Base Feature Selectors						
	NB GLM	**DNN**		**GBM**		**DRF**	
Features	***Importance (%)***			***Features***	***Imp(%)***	***Features***	***Imp(%)***
TreatCCRT	100.00	100.00	100.00	*Nodes*	100.00	*HPV*	100.00
p63	99.00	99.00	99.00	*Age*	85.48	*TreatCCRT*	99.44
Smoke	95.68	95.68	95.68	*Smoke*	83.50	*Nodes*	95.86
Nodes	92.69	92.69	92.69	*StageIV*	71.26	*GradeG3*	95.65
paTT3	92.36	92.36	92.36	*p63*	66.66	*Drink*	95.10
TreatRT	91.36	91.36	91.36	*TreatCCRT*	62.22	*Smoke*	94.25
Invasion	86.05	86.05	86.05	*PaTT4*	61.67	*PlNN2*	92.67
Age	70.43	70.43	70.43	*Size*	61.33	*Age*	92.63
GradeG3	59.14	59.14	59.14	*PaTT3*	52.12	*TreatRT*	91.42
PaTT2	55.81	55.81	55.81	*PlNN2*	47.53	*Invasion*	85.45
SiteNPC	54.82	54.82	54.82	*HPV*	43.82	*p16*	82.76
GradeG2	52.16	52.16	52.16	*PaTT2*	38.44	*p63*	81.36
HPV	47.84	47.84	47.84	*PlNN3*	36.62	*StageIV*	81.05
StageII	47.51	47.51	47.51	*Invasion*	34.40	*PaTT4*	80.91
Drink	39.53	39.53	39.53	*GradeG2*	30.79	*Size*	75.03
SiteOPC	36.21	36.21	36.21	*GradeG3*	29.94	*Gender*	66.81
StageIV	34.88	34.88	34.88	*Gender*	28.51	*GradeG2*	65.34
PlNN3	28.57	28.57	28.57	*TreatRT*	26.44	*SiteNPC*	62.77
Size	27.91	27.91	27.91	*p16*	21.70	*PaTT1*	57.10
PlNN1	27.24	27.24	27.24	*SiteNPC*	15.83	*PaTT3*	54.73

Table 8 shows the top 20 most important features (out of 35) by FS methods. The base feature selectors were learned in an ensemble on the overall dataset, each of which ranked features according to their importance to the class label. Out of 35 features, and by default, each base feature selector ranked the top most 20 features considered important and ignored the rest 15 in this case. To obtain feature subset for heterogeneous ensemble feature selection (HETR-EFS), the top 20 features from all base selectors were aggregated in an ensemble. Features whose ranking range between 80% and 100% were potentially considered important and so, can to be included in the feature subsets summation for optimal feature subset. Then, to determine the optimal feature subset, based on the feature importance in the feature subsets summation, the frequency of each feature was divided by the number of base feature selectors being five (5). Therefore, the feature was considered important or significant in the feature subsets summation if its ratio was at least 0.80 (80%). Table 6 shows the features considered important by each base selector as well as in their ensemble. Based on these feature subsets, the classification models were learned to measure the robustness and effectiveness of each ensemble feature selection technique as well as each stacked-ensemble technique using accuracy, logloss, recall, specificity, and AUC as evaluation metrics.

Table 9. Feature Subset Selected

Features	**Frequency (F)**	**Importance**
Smoke	5	***1.00***
Nodes	5	***1.00***
p63	4	***0.80***
TreatCCRT	4	***0.80***
TreatRT	4	***0.80***
Invasion	4	***0.80***

Table 10. Optimal feature Subsets by Various Ensemble Feature Selection Techniques

EFS Technique	**Feature Subset Selected**
GBM-EFS	*Nodes, Age, Smoke,*
DRF-EFS	*HPV, TreatCCRT, Nodes, GradeG3, Drink, Smoke, PlNN2, Age, TreatRT, Invasion, p16, p63, StageIV, PaTT4*
HETR-EFS	*Smoke, Nodes, TreatCCRT, TreatRT, Invasion, p63*

Table 10 shows the optimal features obtained by various EFS techniques; and these features are: smoking habit (*Smoke*), cervical lymph/neck nodes (*Nodes*), treatment with concurrent chemoradiotherapy (*TreatCCRT)*, treatment with radiotherapy (*TreatRT*), depth of invasion font (*Invasion*), and *p63* type as the most accurate prognosis for HNSCC recurrence based on the available HNSCC dataset.

Table 11. Performance of Stacked Ensemble Model (Model-GBM2) consisting of Two Base Classifiers (GBM and DRF) on Test Data

Ensemble Feature Selection Techniques			
	Heterogeneous	**Homogeneous**	
Metrics	**HETR-EFS**	**GBM-EFS**	**DRF-EFS**
Accuracy	0.8602	0.8172	0.7813
Logloss	0.3014	0.3379	0.5879
Recall	0.6897	0.8939	0.8636
Specificity	0.9375	0.6296	0.6000
AUC	0.8693	0.8018	0.7391

Table 12. Performance of Stacked Ensemble Model (Model-GBM3) Consisting of Three Base Classifiers (GBM, DRF, and DNN) on Test Data

Ensemble Feature Selection Techniques			
	Heterogeneous	**Homogeneous**	
Metrics	**HETR-EFS**	**GBM-EFS**	**DRF-EFS**
Accuracy	0.8925	0.8278	0.7813
Logloss	0.3009	0.3267	0.4167
Recall	0.7333	0.9206	0.8333
Specificity	0.9683	0.6333	0.6250
AUC	0.9164	0.8625	0.8623

Table 13. Performance of Stacked Ensemble Model (Model-GBM4) consisting of Four Base Classifiers (GBM, DRF, DNN, and GLM) on Test Data

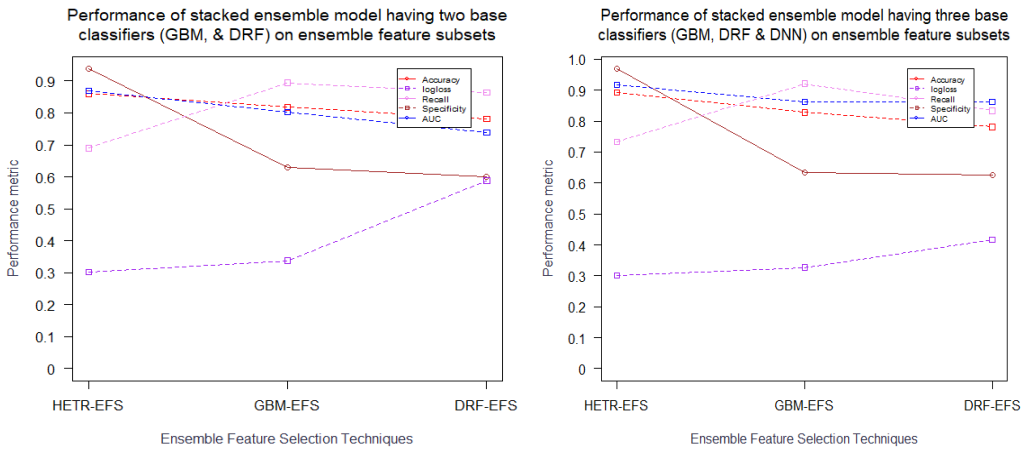
Ensemble Feature Selection Techniques			
	Heterogeneous	**Homogeneous**	
Metrics	**HETR-EFS**	**GBM-EFS**	**DRF-EFS**
Accuracy	0.9032	0.8817	0.8438
Logloss	0.2993	0.3042	0.4141
Recall	0.8261	0.9143	0.6667
Specificity	0.9286	0.7826	0.9500
AUC	0.9058	0.8809	0.8179

Table 14. Performance of Stacked Ensemble Model (Model-GBM5) consisting of Five Base Classifiers (GBM, DRF, DNN, GLM, and NB) on Test Data

Ensemble Feature Selection Techniques			
	Heterogeneous	**Homogeneous**	
Metrics	**HETR-EFS**	**GBM-EFS**	**DRF-EFS**
Accuracy	0.9355	0.9063	0.8817

Logloss	0.2038	0.2959	0.3041
Recall	0.9091	0.7500	0.9265
Specificity	0.9437	1.0000	0.7600
AUC	0.9671	0.9251	0.8321

Considering the Tables 11, 12, 13, and 14, for the data subsets used in this study, best results were obtained using stacked ensemble learning. For the stacked ensemble having two base classifiers on various test data in Table 7, best accuracy (86.02%), log loss (0.3014), specificity (93.75%), and AUC (0.8693) are obtained for data subset provided by heterogeneous ensemble feature selection (HETR-EFS) technique. The best recall (89.39%) is obtained for GBM-EFS data subset. For stacked ensemble model having three base classifiers on test data in Table 8, best accuracy (89.25%), log loss (0.3009), specificity (96.83%), and AUC (0.9164) are obtained for data subset provided by HETR-EFS technique. The best recall (92.06%) is obtained for HETR-EFS data subset. For stacked ensemble model having four base classifiers on test data in Table 9, best accuracy (90.32%), log loss (0.2993), and AUC (0.9058) are obtained for HETR-EFS data subset. The best specificity (95.00%) and recall (91.43%) are obtained for DRF-EFS and GBM-EFS data subsets respectively, the homogeneous ensemble feature selection techniques. For stacked ensemble model having five base classifiers on test data in Table 10, the best accuracy (93.55%), log loss (0.2038), and AUC (0.9671) are obtained for HETR-EFS data subset. Best specificity (100%) and recall (92.65%) are obtained for GBM-EFS and DRF-EFS feature subsets respectively. The graphs of the information in Tables 7, 8, 9, and 10, are represented in Figure 1.



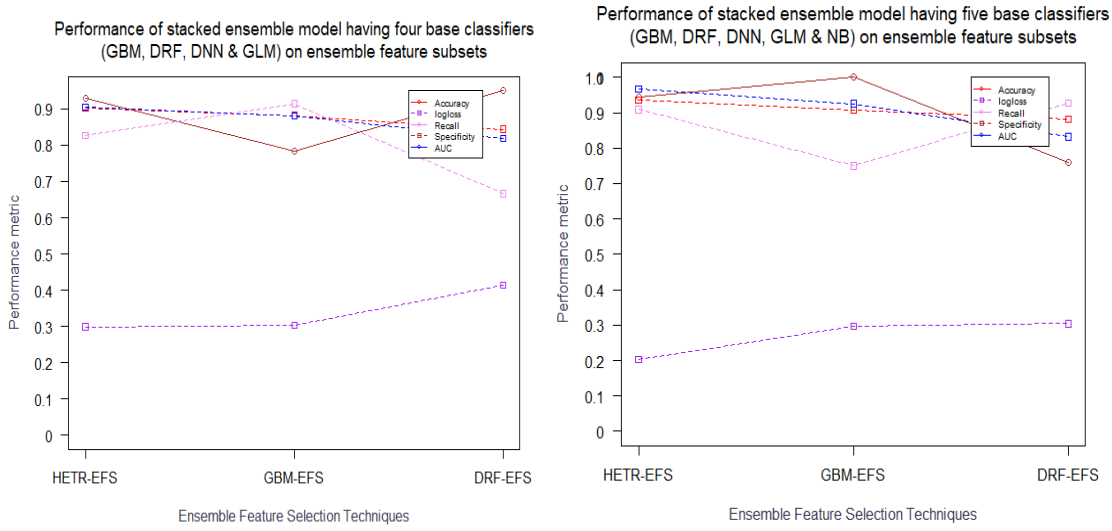


Figure 1: Performance Plots of Ensemble Feature Selection Techniques Based on Stacked Ensemble Techniques

Table 15. Performance comparison of stacked ensemble models on HETR-EFS test set

Metrics	Stacked Ensemble Model-GBM2	Stacked Ensemble Model-GBM3	Stacked Ensemble Model-GBM4	Stacked Ensemble Model-GBM5
Accuracy	0.8602	0.8925	0.9032	0.9355
Logloss	0.3014	0.3009	0.2993	0.2038
Recall	0.6897	0.7333	0.8261	0.9091
Specificity	0.9375	0.9683	0.9286	0.9437
AUC	0.8693	0.9164	0.9058	0.9671

Table 16. Performance comparison of stacked ensemble models on GBM-EFS Test - Set

Metrics	Stacked Ensemble Model-GBM2	Stacked Ensemble Model-GBM3	Stacked Ensemble Model-GBM4	Stacked Ensemble Model-GBM5
Accuracy	0.8172	0.8278	0.8817	0.9063
Logloss	0.3379	0.3267	0.3042	0.2959
Recall	0.8939	0.9206	0.9143	0.7500
Specificity	0.6296	0.6333	0.7826	1.0000
AUC	0.8018	0.8625	0.8809	0.9251

Table 17. Performance Comparison of Stacked Ensemble Models on DRF-EFS Test Set

Metrics	Stacked Ensemble Model-GBM2	Stacked Ensemble Model-GBM3	Stacked Ensemble Model-GBM4	Stacked Ensemble Model-GBM5
Accuracy	0.7813	0.7813	0.8438	0.8817
Logloss	0.5879	0.4167	0.4141	0.3041
Recall	0.8636	0.8333	0.6667	0.9265
Specificity	0.6000	0.6250	0.9500	0.7600
AUC	0.7391	0.8623	0.8179	0.8321

In addition, Tables 15, 16, and 17 show the performance comparison of various stacked ensemble techniques implemented on each ensemble feature subset of the data used in this study. For data subsets provided by each ensemble feature selection technique, the best results are obtained using stacked ensemble learning. Table 15 shows the performance comparison of various stacked ensemble techniques implemented on the test set of HETR-EFS data subset. It can be observed that stacked ensemble technique having five base classifiers performed better than other techniques implemented on the same HETR-EFS subset of the data used in this study. For this data subset, the best accuracy (93.55%), log loss (0.2038), recall (90.91%), specificity (94.37%), and AUC (0.9671) are obtained using stacked ensemble technique having five base classifiers followed by stacked ensemble technique having four base classifiers with accuracy (90.32%), logloss (0.2993), and recall (82.61%). The best specificity (96.83%) is obtained for stacked ensemble technique having three base classifiers.

In Table 16, the best accuracy (90.63%), log loss (0.2959), specificity (100%), and AUC (0.9251) are obtained using stacked ensemble technique having five base classifiers followed by stacked ensemble technique having four base classifiers with accuracy (88.17%) and log loss (0.3042) for GBM-EFS feature subset of the data. The best recall (92.06) is obtained for stacked ensemble technique having three base classifiers. For DRF-EFS subset data, the best accuracy (88.17%), recall (92.65) with the highest log loss (0.3041) are obtained using the stacked ensemble technique consisting of five base classifiers followed by the stacked ensemble technique having four base classifiers; accuracy (84.38%) with the log loss (0.4141), and the specificity (95.00%). The best AUC (86.23%) is obtained for stacked ensemble technique having three base classifiers. The graphs of the information in Tables 11, 12 and 13 are represented in Figure 2

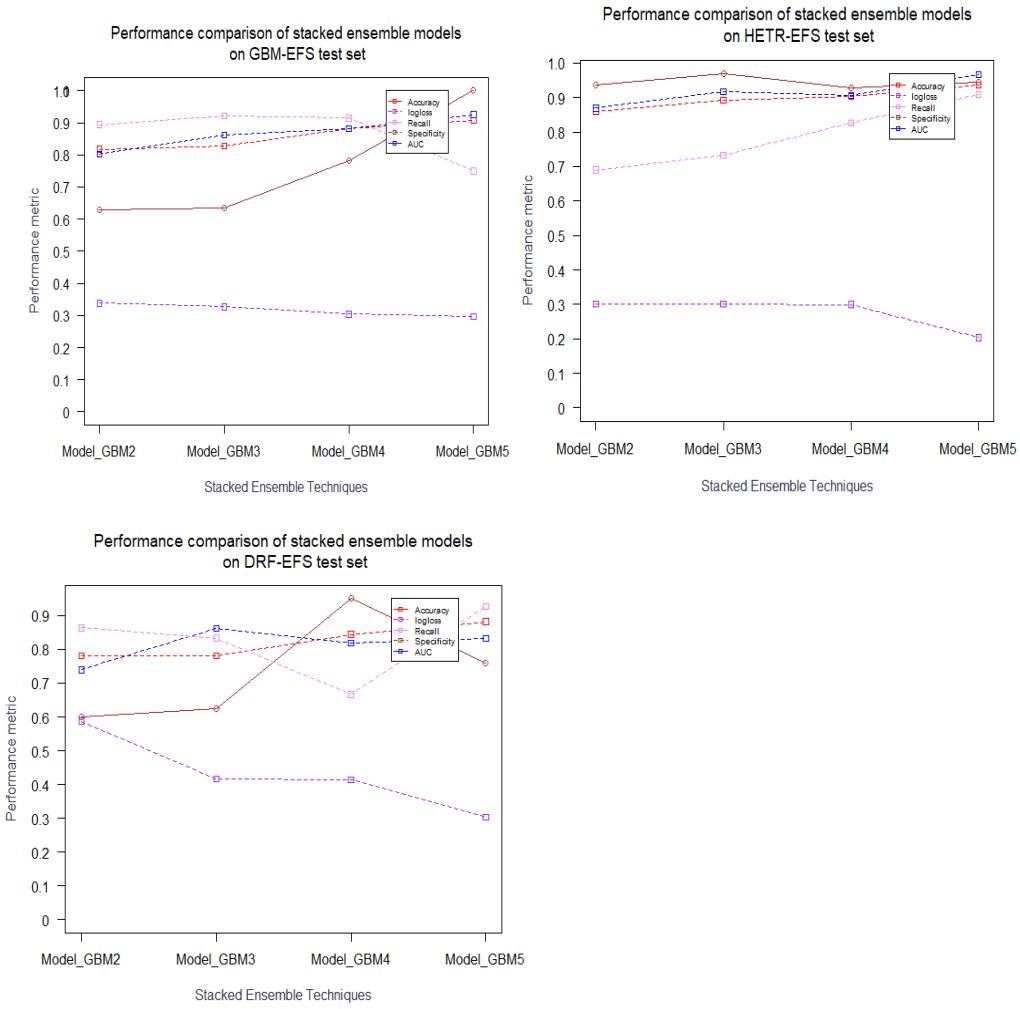
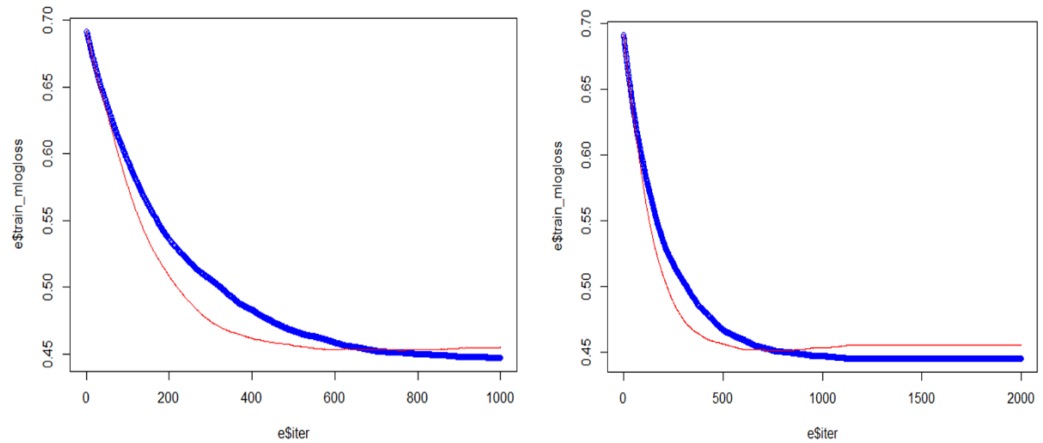


Figure 2: Performance plots of Stacked ensemble techniques on various ensemble feature subsets



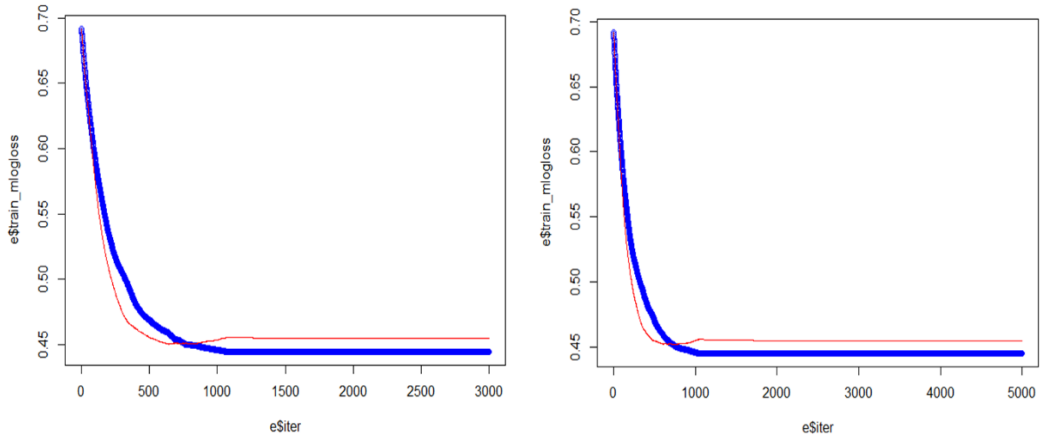


Figure 3. A Plot of Good Fit Learning Curves

Figure 3 displays the stacked ensemble model with five base classifier models' good fit learning curves graphic. The training loss (blue curve) and validation loss (red curve) of a model should both decline to a level of stability and flatten at the point when they can no longer decrease. It was found that both curves descended to a stable point with a narrow separation between them known as the *generalisation gap*. This demonstrates that adding training examples does not enhance a model's performance when there has been a training loss and that adding training examples does not enhance a model's performance when there has been a validation loss. This demonstrated how well the suggested stacked ensemble model with five base classifiers suited the data.

In summary, the results of various stacked ensemble techniques implemented on feature subsets of the data provided by various ensemble feature selection techniques used in this study showed that, all the stacked ensemble techniques used in this study achieved higher performance on data subset provided by heterogeneous ensemble feature selection technique compared to their performance results on data subsets provided by other ensemble feature selection techniques. It was also observed that for each stacked ensemble technique implemented on each feature subset of the data provided by ensemble feature selection techniques, the stacked ensemble technique consisting of five base classifiers achieved the highest accuracy coupled with the least log loss compared to other stacked ensemble techniques used in this study. In terms of AUC, it was also observed that for each stacked ensemble technique implemented on each feature subset of the data provided by ensemble feature selection techniques with the exception of DRF feature subset data, the stacked ensemble technique consisting of five base classifiers achieved the highest AUC compared to other stacked ensemble techniques used in this study. The stacked ensemble technique having three base classifiers achieved the highest AUC on data subset provided by DRF feature selection technique.

Even though the individual ensemble feature selectors GBM and DRF, performed well under various stacked ensemble models, the performance

improved across board when the they were combined with single feature selectors DNN, GLM, and NB to form the heterogeneous ensemble features. The stacking ensemble utilising the heterogeneous ensemble features, on the other hand, demonstrated superior prediction accuracy than the pre-existing base ensemble models into consideration as ensemble feature selection techniques in this study. The outcomes of this study demonstrated that stacked ensemble models with heterogeneous ensemble features are useful as a supplementary tool for categorising and predicting recurrent HNSCC prognostic data.

CONCLUSIONS AND RECOMMENDATIONS

This study presented a stack-ensemble model by combining five (5) ML algorithms such that all were adopted as base-classifiers and as specific model as the meta-classifier using the H2O package in R programming language which is an open-source library from H2O.ai. This paper focused on the improvement of the ensemble classification performance through stacked generalisation and feature selection toward the prediction of HNSCC recurrence patterns using data subsets provided by various EFS techniques considered in this study. To achieve this, the SE technique that finds the optimal weighted average of diverse machine learning base classifier models using meta-learning algorithm was used. For base classifiers, GBM and DRF were used and another base classifier DNN along with the previous two (GBM and DRF) was integrated. Next, another base classifier GLM along with the previous three (GBM, DRF, and DNN) was integrated. Then, another base classifier NB along with the previous four (GBM, DRF, DNN, and GLM) was integrated. To achieve the optimal combination of these diverse base classifier models with EFS technique considered in this study, the GBM was used as meta-classifier based on its high performance when trained as a base classifier model on various ensemble data subsets compared to other base classifier models considered in this study. The experimental results showed that using stacked ensemble technique having five base classifiers had better performance compared to other stacked ensemble techniques considered in this study, and using heterogeneous ensemble feature selection technique is better as a supporting tool for generating the most accurate prognostic features for HNSCC dataset.

ADVANCED RESEARCH

Each study has limitations; thus, you can describe it here and briefly provide suggestions for further research.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. Joel Yarney, the Head of National Centre for Radiotherapy and Nuclear Medicine, Dept. of Radiotherapy and Oncology at Korle Bu Teaching Hospital (KBTH), Accra, Ghana; Mr. Philip,

and Mr. Charles at the Cancer Registry at KBTH for their support in data collection for the conduct of this study.

REFERENCES.

Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. Journal of Supercomputing, 73(11), 4773–4795. https://doi.org/10.1007/s11227-017-2046-2

Apalla, Zoi, Lallas, A., Sotiriou, E., Lazaridou, E., Ioannides, D., 2017. Epidemiological trends in skin cancer. Dermatol. Pract. Concept. 7, 1–6. https://doi.org/10.5826/dpc.0702a01

Correia de Sá, T.R., Silva, R., Lopes, J.M., 2015. Basal cell carcinoma of the skin (part 2): diagnosis, prognosis and management. Future Oncol. Lond. Engl. 11, 3023–3038. https://doi.org/10.2217/fon.15.245

Garg, Tanya, and Yogesh Kumar. 2014. "Combinational Feature Selection Approach for Network Intrusion Detection System." 2014 International Conference on Parallel, Distributed and Grid Computing, 82–87. https://doi.org/10.1109/PDGC.2014.7030720.

Yarney et al. (2017), "Cancers of the Head & Neck: Does concurrent chemoradiotherapy preceded by chemotherapy improve survival in locally advanced nasopharyngeal cancer patients? Experience from Ghana", BioMed Central, No. 2, Vol. 4, pp. 1-7