Journal of Mathematical Problems, Equations and Statistics

E-ISSN: 2709-9407
P-ISSN: 2709-9393
JMPES 2023; 4(1): 121-134

www.mathematicaljournal.com
Received: 03-04-2023
Accepted: 04-05-2023

Damianus Kofi Owusu
Ph.D., Department of Mathematical Sciences, Faculty of Engineering, University of Mines and Technology, Tarkwa, Ghana

Peter Kwesi Nyarko
Ph.D., Department of Mathematical Sciences, Faculty of Engineering, University of Mines and Technology, Tarkwa, Ghana

Corresponding Author:
Damianus Kofi Owusu
Ph.D., Department of Mathematical Sciences, Faculty of Engineering, University of Mines and Technology, Tarkwa, Ghana

Stacked ensemble model for recurrent head and neck squamous cell carcinoma prognosis based on clinicopathologic and genomic markers

Damianus Kofi Owusu and Peter Kwesi Nyarko

Abstract
The prevalence of head and neck squamous cell carcinoma (HNSCC) and its recurrences is not declining in Ghana as a result of the disease's delayed diagnosis and dismal prognosis. Early detection and treatment are crucial since HNSCC recurrence and tumor stage at diagnosis are significantly correlated. This study looked at the best meta-classifier model where the same ML classifiers for base classifiers and meta classifiers are employed in order to determine the most reliable prediction and robust prognostic model for recurrent HNSCC. Based on gradient boosted features (GBF), the suggested model was an ensemble of ML models that were stacked. Each of these models served as a meta-classifier and as a building block for the base classifiers. To find the optimal meta-classifier model, the performances of different meta-models were compared. The findings demonstrated that utilising the GBM as a meta-classifier produced superior accuracy with the least log loss compared to that produced by any other model of recurrent HNSCC prognostic data. This gave a stacked ensemble model termed as a HESCA model, consisted of five base models and GBM meta-model. 8-input HESCA model was compared with full-input model, and 8-input HESCA model was also compared with 8-input models. The results of the study demonstrated that using a GBM classifier as a meta-classifier in a stacking ensemble with five base classifiers based on GBF or GBM input features outperformed standalone models and any full-input model. Additionally, using a GBM as a meta-classifier is appropriate as a supporting tool for identifying, classifying, and predicting recurrent HNSCC prognosis data.

Keywords: Recurrent HNSCC prognosis, ensemble learning, stacked ensemble, classification

Introduction
According to [8], the number of cancer cases, particularly Head and Neck Squamous Cell Carcinomas (HNSCCs) and their recurrences, is not declining in Ghana as a result of people delaying their visits to medical institutions until they exhibit cancer signs and symptoms. Most of these patients turn to have tumors at the advanced or metastatic stage at diagnosis. Based on the research by [20] that recurrent cancer is strongly linked to the stage of the tumor at diagnosis, most of these patients have about 60-90% probability of experiencing cancer recurrence even after a successful treatment where cancer had reached its remission. That is, early diagnosis and treatment can help reduce cancer recurrence by identifying accurate prognostic markers [8, 6, 22]. 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 various fields using Statistics, Artificial Intelligence (AI), and ML techniques, attempting to address the challenge of the patient's disease recurrence [5, 10]. They include: [22] used the Kaplan-Meier analysis and the Log-rank test to determine if patients at KBTH would survive nasopharyngeal carcinoma. Using simple descriptive analysis using the International Classification of Diseases coding system [16], reclassified sociodemographic, clinical, and pathological data on patients with HNC at KATH. At KBTH [1], reclassified patients with oral cavity and oropharynx Squamous Cell Carcinoma (SCC) using the ICD-10, the 10th revision of the International Statistical Classification of Diseases and Related Health Problems. In order to estimate the number of cases [14], additionally examined epidemiological (clinical and histopathologic) characteristics for laryngeal cancer in SCC patients at KBTH [9]. Used AI/ML algorithms to categorize tumors as malignant or benign in a study on breast cancer. When attempting to forecast the labels of upcoming, unobserved data, statistical and standalone ML approaches fall short [4, 2]. Finding a classifier that performs well when predicting the labels of future, ambiguous data is the aim of classification [21].

Usually, in cancer diagnosis and prognosis; clinical data, pathological data, genomic data, and socio-demographic data are integrated in the prognostic model for robust prognosis [6, 9]. Now, given the rapid development of these medical information and the growing trend and reliance on the application of machine learning techniques in cancer domain, it is worthwhile that if these information on patients are combined [7, 10] under some variant studies such as ensemble ML techniques [4, 2], more accurate prognosis can be generated for early diagnosis and treatment outcomes [15, 17]. A single classification model that has good generalization ability is difficult to qualify as a strong classifier, but ensemble learning can turn a collection of weak classifiers into a strong one by combining them to produce one with good generalization ability [21]. A decent generalized classification model should be broad enough to account for cases that haven't been observed before rather than overfitting the training set. This way, to achieve a good generalization ability of a classification model, stacking or stacked generalization [4, 2, 12, 13, 21] with feature selection, is the best way to go. There are very few studies that used stacking ensemble techniques in the prognosis of HNSCC.

Also, in order to obtain the most promising prognostic classification model, one needs to combine single base classifiers and weak ensemble classifiers. Currently, many studies on HNC had focused on stacking ensemble techniques having only single base classifiers (without integrating standalone ensemble classifiers) in the domain of various cancer studies. To classify the prognosis of recurrent breast cancer, for example [2], constructed a stacked ensemble-based model (with 10-fold CV) using two single base classifiers in combination of DT and SVM, NB and SVM, and NB and DT, where DT, NB, and SVM were employed as meta-classifiers, respectively [4]. Developed a stacked ensemble model with three single base classifiers, KNN, NB, and DT (C4.5), and GLM as a meta-learner, able to predict the cancer kinds in the vicinity of the HNC regions (Sinonasal, nasopharyngeal, laryngeal, and thyroid). In order to promote quick referral [3], used the same technique (KNN, NB, and DT (C4.5) as base learners, and GLM as a meta-learner) to diagnose HNC susceptibility. These studies' prognostic models might not have strong generalizability for predicting the prognosis for HNC. While bagging, boosting, and stacking are three major meta-algorithms that offer effective methods of combining base learners, stacking is the most effective method, especially when it combines single base learners and standalone ensemble learners. This has been demonstrated by [12] in various healthcare datasets (Wisconsin Breast Cancer, Pima Indian Diabetes Dataset, Indian Liver Patient Dataset, [13] in breast cancer, and) [21]. In Ghana, based on the literature, there has not been any study yet on any form of stacked ensemble ML algorithms on genomic and clinicopathologic makers for recurrent HNSCC prognosis [8] that is prone to provide unbiased, stable, and reliable prognostic outcomes [22]. Therefore, due to this medical gap in cancer predictive foci, this study seeks to investigate how stacked ensemble ML techniques can be employed in the recurrent HNSCC prognosis in the developing country Ghana, to address the issue of poor and contradictory prognosis produced by biased, unstable, and unreliable prognostic models in existence; in order to identify, classify and prognosticate the most stable and accurate prognosis for recurrent HNSCC patients, being the first study ever. This method is an organized effort to develop a Hybrid Ensemble Super Classification Algorithm (HESCA) prognostic model using a stacked generalization of various supervised machine learning techniques with an ensemble feature selection method in Ghana. In addition to developing a hybrid stacked ensemble-based classification model that combines the best set of multiple base classifiers and the GBM feature selection method for the prognosis of HNSCC recurrence, a multifaceted complementary approach to the study's design is likely to simultaneously capture the most accurate prognosis.

The current study, which, according to the reviewed literature, is the first study in Ghana, differs from the aforementioned studies in that it focuses on the prognosis of recurrent HNSCC using a stacked ensemble technique with five base classifiers (GBM, DRF, DNN, GLM, and NB) with GBM feature selection. The homogenous ensemble classifiers (GBM and DRF), and standalone single classifiers (DNN, GLM, and NB) as proposed by [13, 12] to be the most effective algorithms for stacked generalization for the classification and prediction of cancer cases, have been employed under this study. Furthermore, the system is tested using the data collected locally at KBTH; the National Centre for Radiotherapy and Nuclear Medicine (NCRNM), Radiotherapy and Oncology department. In order to compare the findings, the same proposed HESCA model was trained on both the original dataset and the best feature subset produced by the GBM feature selection technique. In addition, the results generated from the proposed HESCA model are compared with standalone models' results using the same dataset provided by the GBM feature selection technique. Lastly, the proposed HESCA model is validated using the existing HNSCC test data.

To this end, due to ever-increasing in the recurrent rate of HNSCC in the developing country, there is a need to develop a more robust computerized tool that is needed to aid clinicians in the decision support stage and to identify the most accurate prognosis so as to better prognosticate the rate of recurrence for each HNSCC patient and to extend the model to other cancer prognosis prediction for early diagnosis and treatment. The overall goal of the study is to create a stacked ensemble classification model that combines standalone ensemble classifiers and single base classifiers to provide a robust prognosis for early diagnosis and treatment outcomes based on the best feature subset of clinical, histopathologic (pathologic), and genomic markers, as well as other risk factors and treatment options related to HNSCC recurrence in Ghana. This is to improve early diagnosis and primary treatment of the malignant tumor that minimises its recurrence after it reaches remission.

Ensemble Learning

The term "ensemble learning" refers to a group (or ensemble) of base learners or models that collaborate to provide a reliable final prediction. Because of significant variation or high bias, an individual learner, sometimes referred to as a base or weak learner, might not do effectively on their own. The aggregation of weak learners, however, might result in the formation of a strong learner since it decreases biases or variances, which improves classifier performance [21]. Decision trees are frequently used to illustrate ensemble techniques, but they can be vulnerable to underfitting (low variance and high bias) when they are very small, such as decision stumps, which are decision trees with only one level, and overfitting (High variance and low bias) when pruning has not been done. It is important to note that a learner cannot generalize well to new or unexplored datasets when the training data is either overfitted or underfitted. Ensemble

approaches are used to prevent this behavior and enable the learner to generalize to fresh training samples. Although decision trees can show strong bias or high variance, it is important to note that other modeling techniques also use ensemble learning to identify the "sweet spot" in the tradeoff between bias and variance. When learning the ensemble model, two primary methods are taken into account: Homogeneous ensemble learning, of which classifiers pool the predictions of multiple individual decision trees. This category of ensemble learning is broadly highlighted in two techniques; bagging and boosting. The second method, called heterogeneous ensemble learning, uses various basic classifier types to build a heterogeneous ensemble model that is employed in stacking. Generally speaking, three main meta-algorithms offer efficient methods for integrating weak or base learners [18]:

Bagging

Bagging is also known as Bootstrap Aggregation, which is an ensemble ML technique that combines homogenous base learners into a more robust learner. With replacement idea of drawing dataset at random, and using these different random subsets of the data to train different classifiers is what is called bootstrapping. If this technique is used to combine individual classifiers (decision trees), this process is called bagging. So, bagging simply means, constructing each classifier or tree on a different random subset of the dataset drawn with replacement. The predictions from each independent classifier can be combined by averaging (Regression) or by majority voting (classification) to derive the final prediction. A widely used algorithm or technique is Random Forest (RF).

Boosting

Boosting is a homogeneous learning strategy that creates a strong learner out of a homogenous group of weak (decision tree) learners in order to reduce training error. In boosting, a randomly chosen sample of data is chosen, fitted with the learner, and then consecutively learned. In other words, each learner 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 methods of adaptive, gradient, and extreme gradient boosting (AdaBoost, Gradient Boost, and XGBoost) are the main emphasis of the strategies for boosting [18]. For the purpose of the study, Gradient Boost or GBM is discussed and used for feature selection.

Stacking

Stacking is a technique that combines heterogeneous multiple base learners into a more robust learner in their combination. This technique combines the predictions made by different individual learners into make a final bust prediction. A meta-learner with less variance and bias can be created when base learning algorithms are properly integrated [18]. Cross-validation is used by stacking to gauge the effectiveness of various base learning methods [11]. The meta-learning algorithm (s) are fed with the output from the base learners, also known as "meta-features" in the stacking literature [19]. A high-level classifier is learned through stacking on top of the base classifiers. It can be viewed as a meta-learning strategy in which the first-level classifiers, which serve as the foundational classifiers, are combined to learn a second-level classifier known as a meta-classifier [18].

Dataset and evaluation metrics

Data Source

A retrospective cohort analysis of 125 HNSCC patients under the age of 15 who had previously been diagnosed with the disease, received curative treatment at KBTH and were monitored until the cancer had gone into remission but then between 2016 and 2020 either experienced recurrence or nonrecurrence. Information on each patient's gender, age at diagnosis, alcohol consumption, smoking habits, habit of chewing tobacco, primary tumor site, tumor stage at diagnosis, histological grade, and tumor size is also included. invading the front's depth. The following factors are taken into account: cervical lymph nodes, pathological tumor staging, pathological lymph nodes, family history of cancer, human papillomavirus level, p16 type, p63 type, and kind of treatment. The features evaluated in this study are listed in Table 1.

Table 1: Demographic, Clinicopathologic, and Genomic Features

	Feature Name	Description
1	Gen	Gender
2	Age	Age at diagnosis
3	Alc	Alcohol drinking habit
4	Smoke	Smoking habit
5	Chew	Quid/Tobacco chewing habit
6	Site	Primary site of tumor
7	Stage	Tumor stage at diagnosis
8	Grade	Histological grade
9	Size	Tumor size
10	Inv	Depth of invasion front
11	Nodes	Cervical lymph/Neck nodes
12	PaT	Pathological tumor staging
13	PlN	Pathological lymph nodes
14	FHx	Family history of cancer
15	HPV	Human papillomavirus type
16	p16	p16 type
17	p63	p63 type
18	Treat	Treatment type

Data pre-processing

A normalized predictive mode approach was used to identify and fill in the missing examples using mode imputation. This technique is feasible in this study as the size of training examples is very small in order to avoid the deletion of the instances having missing training examples under any feature. To provide a normalized dataset for training, evaluation, and prediction, one-hot encoding was employed for features with more than two levels during data discretization and transformation. As a result, the dataset's 35 features instead of the original 18 characteristics were to be taken into account for model learning. Table 2 describes features categorization into levels that were ready for ingestion into the process of feature selection prior to model training and evaluation.

Table 2: Description of Features for 125 Instances

Feature	Description	Levels	Feature	Description	Levels
Gen (x_1)	Male Female	0 1	Inv (x_{10})	Cohesive Non-cohesive	0 1
Age (x_2)	15-45 > 45	0 1	Node (x_{11})	Positive Negative	0 1
Alc (x_3)	Yes No	0 1	PaT (x_{12})	T_1 T_2 T_3 T_4	0 1 2 3
Smoke (x_4)	Yes No	0 1	PlN (x_{13})	N0 N1 N2 (N2a, N2b, N2c) N3 (N3a, N3b)	0 1 2 3
Chew (x_5)	Yes No	0 1	FHx (x_{14})	Yes No	0 1
Site (x_6)	Larynx Nasopharynx Oropharynx Hypopharynx	0 1 2 3	HPV (x_{15})	High-risk (HPV16&18) Low-risk (HPV6&11)	0 1
Stage (x_7)	I II III IV	0 1 2 3	p16 (x_{16})	Positive Negative	0 1
Grade (x_8)	G1 G2 G3	0 1 2	p63 (x_{17})	Positive Negative	0 1
Size (x_9)	0-4cm > 4cm	0 1	Treat (x_{18})	Surgery only Radiotherapy (RT) only Chemotherapy (Chemo) only Concurrent ChemoRT (CCRT)	0 1 2 3

NB: Cohesive if depth of invasion ≤10mm, and Non-cohesive if depth of invasion >10mm. G1: Welldifferentiated, G2: Moderately differentiated, G3: Poorly differentiated, G4: Undifferentiated

Performance metrics

The performance metrics of a classification model based on GBF of recurrent HNSCC prognosis dataset, that are mostly used in cancer study to assess the model performance were utilized. These metrics are; accuracy, recall, precision, specificity, F1-score, and Area Under Receiver Operating Characteristic Curve (AUROC), including the logarithmic loss.

Table 3: Confusion matrix for recurrent HNSCC prognosis

		Actual conditions	
		Recurrence (+)	**Nonrecurrence (-)**
Predicted outcomes	Recurrence (+)	TP	FP
	Nonrecurrence (-)	FN	TN

True positive (TP), False positive (FP), False negative (FN), True negative (TN)

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

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \times 100\%$$

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP+FP}} \times 100\%$$

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

$$\text{F1} - \text{score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Binary Cross-Entropy/Logarithmic Loss

The logarithmic loss (log loss) metric can be used to evaluate the performance of the binomial or multinomial classifier. Log loss is the negative average of the log of corrected predicted probabilities for each instance. It measures the uncertainty of the predicted labels based on how far it varies from the actual label. Log loss equation for binary classification is;

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

where N is the total number of rows (observations) of the corresponding data frame, w is the per row user-defined weight (default is 1), p is the predicted value assigned to a give row (observation)

Proposed HESCA model development

Figure 1 represents the architecture of the proposed HESCA model with full-input features in stacking ensemble. It has 35 input features based on one-hot encoding. Figure 2 on the other hand, represents the architecture of the proposed HESCA model with 8-input features provided by GBFS technique. Base or standalone classifiers were initially trained based on optimal feature subset provided by GBM features. Next, 10-fold cross-validation was performed on each base classifier; each of which provided cross-validated predictions called meta-features as input for meta classifiers. These cross-validated predicted labels along with the original class labels gave the level-one data to learn meta-classifiers. Then, each base classifier was used as meta classifier to learn meta-models base on the level-one dataset. Now, based on the outputs of these meta classifiers, the best meta classifier was identified and its output served as the final prediction. This is the stacked ensemble model consisting of five base models and GBM meta model. Table 4 presents the hyperparameters via random grid search used to learn HESCA model

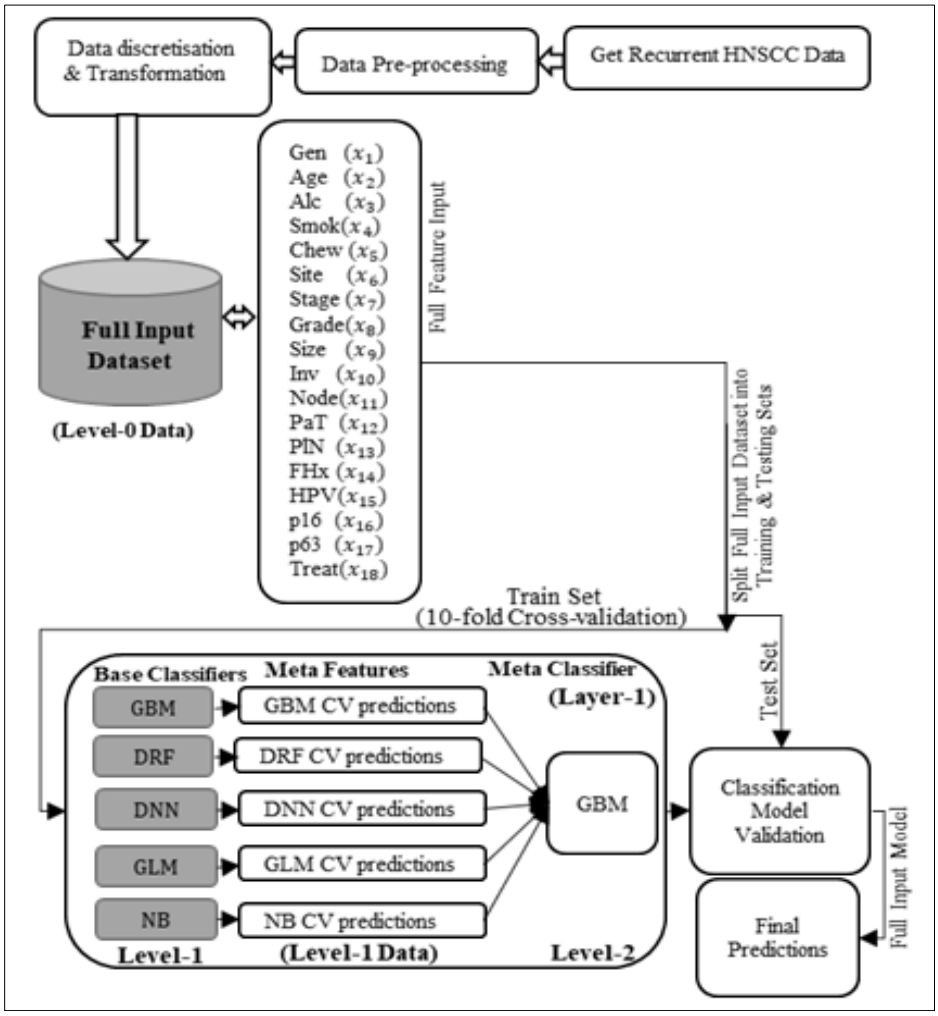


Fig 1: Architecture of HESCA model with full-input features using stacking

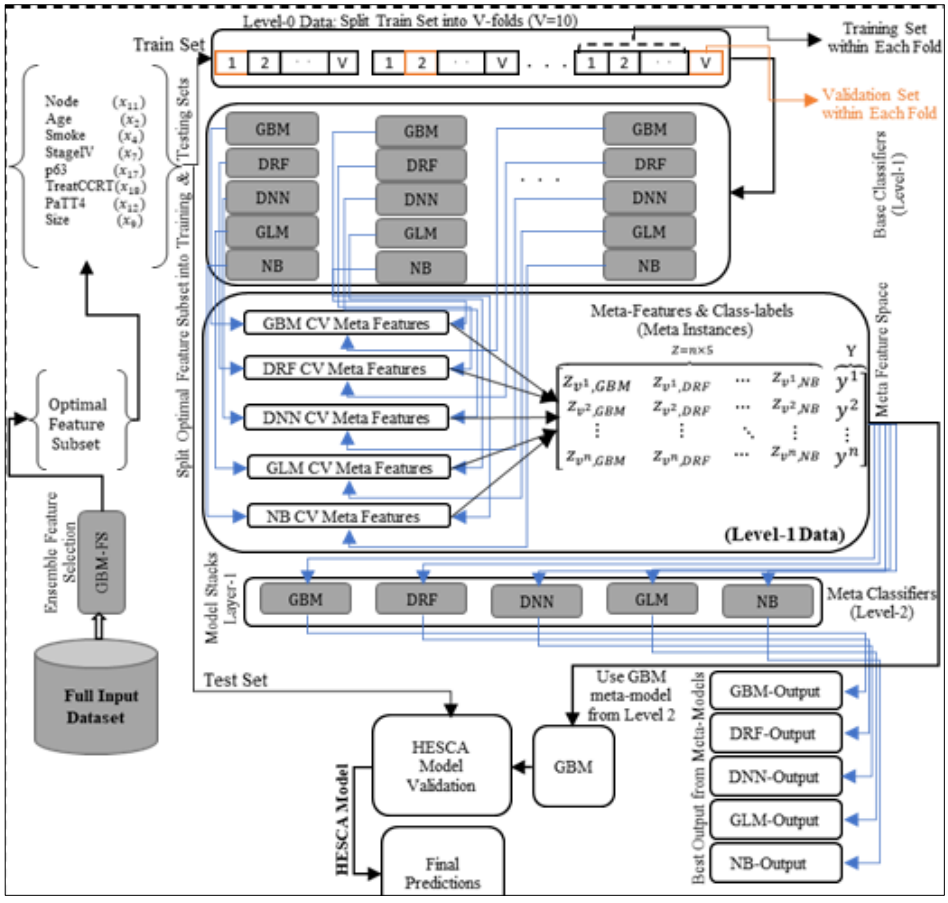


Fig 2: Architecture of HESCA model with 8-input features using stacking

Table 4: 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), Min rows = c (1, 3, 5), Learn rate = c (0.01, 0.1), Learn rate annealing =c (0.99, 1), Sample rate= c (0.5, 0.75, 1), Col sample rate =c (0.8, 0.9, 1)	trees = 5000 unfolds = 10 fold assignment = "Modulo" keep cross validation predictions = TRUE stopping rounds = 50
DRF	Max depth = c (9, 30), entries = 3, min rows = c (1, 5, 10), sample rate = c (0.5, 0.75, 1), col sample rate per tree= (0.8, 0.9, 1)	trees = 5000 unfolds = 10 fold assignment = "Modulo" keep cross-validation predictions = TRUE stopping rounds = 50
DNN	activation=c("Rectifier","Maxout", "Tanh"), hidden = list (c (5, 5, 5, 5), c (10, 10, 10, 10), c (50, 50, 50, 50), epochs = c (50, 100, 200), l1 = c (0, 1e-3, 1e-5), l2 = c (0, 1e-3, 1e-5), rate =c (0, 0.1, 0.005, 0.001), momentum start = c (0, 0.5), momentum stable=c (0.99, 0.5)	epochs = 20 unfolds = 10 fold assignment = "Modulo" keep cross validation predictions = TRUE stopping rounds = 50
NB	Laplace=c(0, 5, by 0.5)	unfolds = 10 fold assignment = "Modulo" keep cross-validation predictions = TRUE
GLM	Alpha=c(0.1)	unfolds = 10 remove collinear columns = TRUE fold assignment = "Modulo" keep cross-validation predictions = TRUE

Evaluation results and discussions

Results

Figure 2 shows that the ranking of features according to their importance in an ensemble by Gradient Boosted Feature Selection (GBFS). A minimum criterion of 60% was taken into account in order to produce the ideal feature subset; hence, features with rankings between 60% and 100% were possibly important. *Nodes*, *Age*, *Smoke*, *Stage IV*, *p63*, *Treat CCRT*, *PaTT4*, and *Size* are among these characteristics.

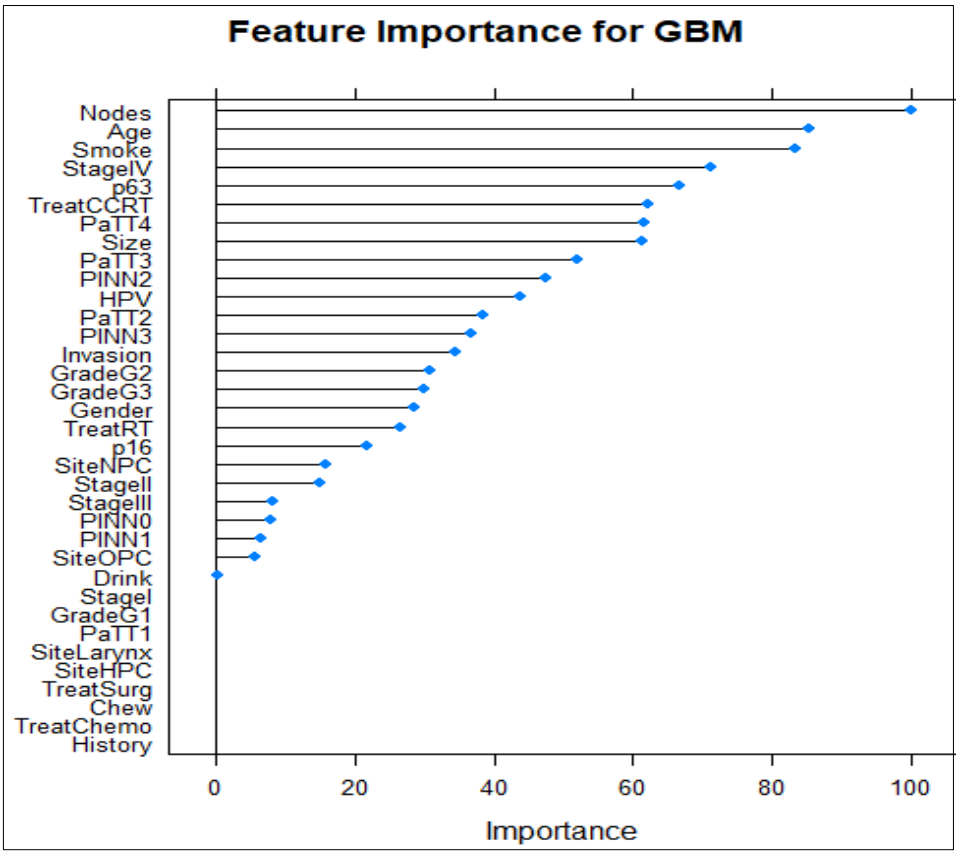


Fig 3: Rank of features by Gradient Boosted Feature Selection (GBFS)

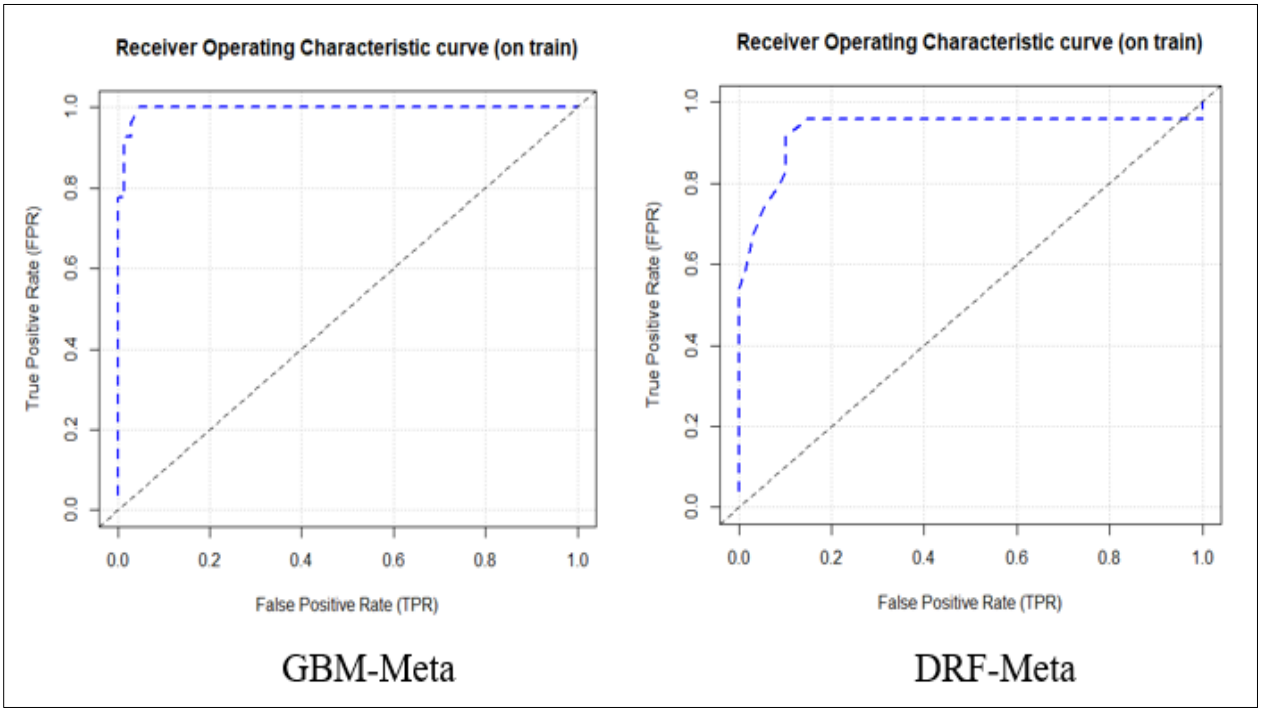
Table 5: Performance metrics of Meta classifiers on level-one training set

	Meta Classifiers				
Metrics	**GBM**	**DRF**	**DNN**	**GLM**	**NB**
Accuracy	0.9677	0.9139	0.9247	0.9355	0.9355
Logloss	0.1172	0.3139	0.5123	0.2986	0.2038
Recall	0.9000	0.8333	0.8400	0.9091	0.9091
Specificity	1.0000	0.9420	0.9559	0.9437	0.9437
Precision	1.0000	0.8333	0.8750	0.8333	0.8333
F1-Score	0.9474	0.8333	0.8571	0.8696	0.8696
AUC	0.9952	0.9134	0.9199	0.9834	0.9671

Table 6: Performance metrics of Meta Classifiers on Test Set

	Meta Classifiers				
Metrics	**GBM**	**DRF**	**DNN**	**GLM**	**NB**
Accuracy	0.9063	0.7813	0.8750	0.7813	0.8750
Logloss	0.2959	0.5095	0.5854	0.4406	0.4208
Recall	0.7500	0.5625	0.8571	0.5714	0.7273
Specificity	1.0000	1.0000	0.8800	0.9444	0.9524
Precision	1.0000	1.0000	0.6667	0.8889	0.8889
F1-Score	0.8571	0.7200	0.7500	0.6957	0.7999
AUC	0.9251	0.7149	0.8937	0.9179	0.8961

Table 5 shows the performance of different training metrics of meta classifiers on the level-one training set considered in this study. The performances metrics of meta classifiers were obtained by learning each base classifier on the cross-validated predicted labels along with the original class labels. Best results are obtained using stacked ensemble techniques. The GBM meta classifier had best accuracy (0.9677), log loss (0.1172), specificity (1.00), precision (1.00), F1-Score (0.9474), and AUC (0.9952) while GLM meta classifier had the best recall value (0.9091). Table 6 shows the evaluation performance of meta classifiers on test set. It can be observed that best results are obtained for GBM meta classifier with the highest accuracy value (0.9063), F1-Score (0.8571) and AUC value (0.9251) and the least log loss metric (0.2959) as compared to that of other meta classifiers. GBM and DNN meta classifiers both had the best specificity and precision metrics (1.00), and DNN meta classifier had the best recall metric (0.8571). The ROC curve analysis of each meta classifier on training set and test set is shown in Figure 4 and 5 respectively.



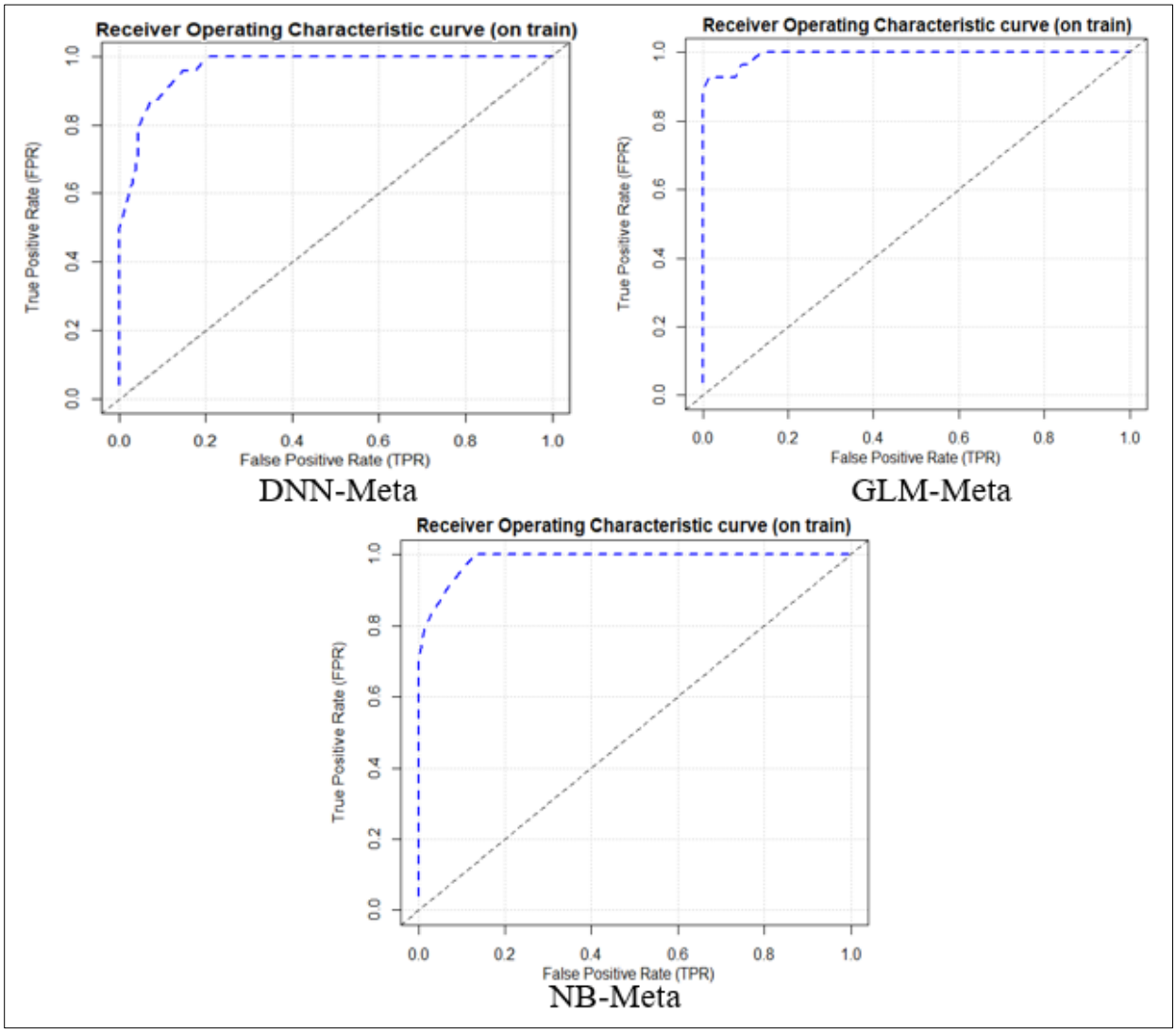
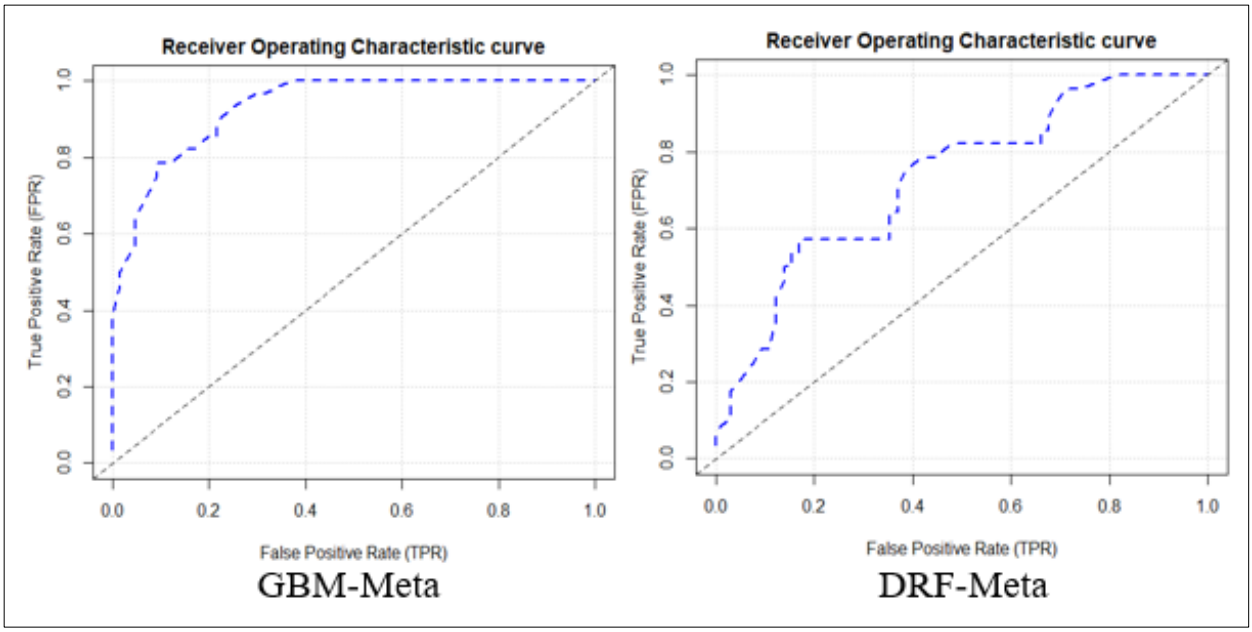


Fig 4: ROC Curve Analysis of Meta Classifiers on Training set



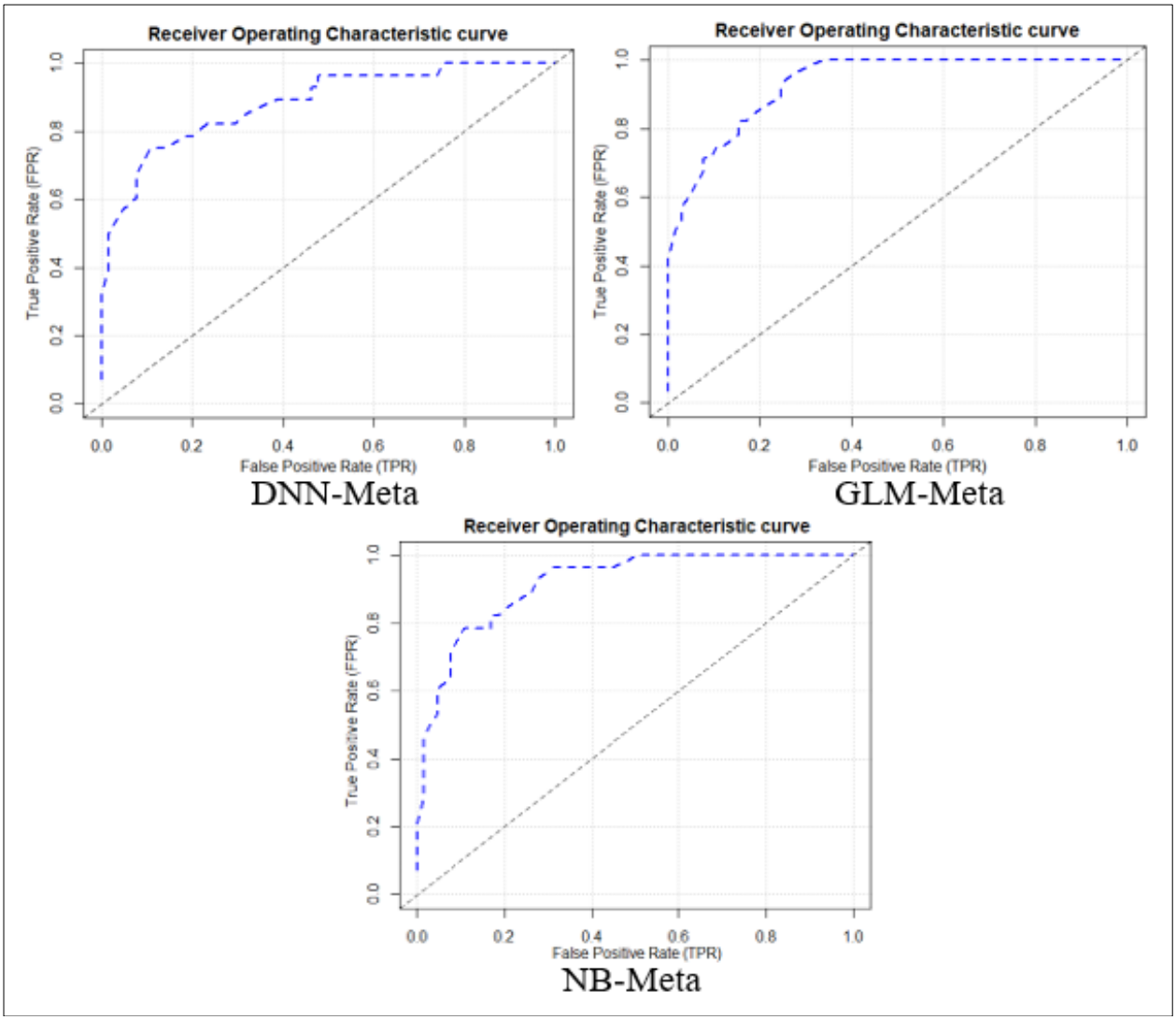


Fig 5: ROC Curve Analysis of Meta Classifiers on Test set

Table 7: Performance Comparison of HESCA model with full-input features and HESCA model with GBF

Metrics	Training set		Testing set	
	HESCA Model with full-input Training Set	HESCA Model with GBF	HESCA Model with full-input Test Set	HESCA Model with GBF
Accuracy	0.3441	0.9677	0.3438	0.9063
Logloss	0.8025	0.1172	1.0435	0.2959
Recall	0.3023	0.9000	0.3846	0.7500
Specificity	0.8571	1.0000	0.1667	1.0000
Precision	0.9629	1.0000	0.6667	1.0000
F1-Score	0.4602	0.9474	0.4878	0.8571
AUC	0.4879	0.9952	0.4364	0.9251

Table 7 shows the performance of the HESCA model with full-input features and the HESCA model with GBF. The performance metrics on the training set were recorded to assess the performance of the model on the training set, and on and test set, to evaluate how well the model will perform on unseen labels. Based on accuracy and log loss, the best accuracy (0.9677) with the log loss (0.1172) and accuracy (0.9063) with the log loss (0.2959) on the training set and test set respectively are obtained for the HESCA model with GBF as compared to the accuracy (0.3441) with the log loss (0.8025) and accuracy (0.3438) with the log loss (1.0435) obtained on the training set and test set respectively for HESCA model with full-input features. AUC (0.9952) and (0.9251) obtained on the training set and test set respectively are for HESCA model with GBF compared to the AUC (0.4879) and (0.4364) obtained on the training set and test set respectively for HESCA model with full-input features. Based on the training set, the HESCA model with GBF had a recall value (0.9000), specificity value (1.00), precision value (1.00) and F1-Score (0.9474) as compared to the HESCA model with full input features that had recall value (0.3023), specificity value (0.8571), precision value (0.9630) and F1-Score (0.4602). Similar to the test set, the HESCA model with GBF had a recall value (0.7500), specificity value (100%), precision value (1.00), and F1-Score (0.8571) compared to the HESCA model with full input features that had the recall value (0.3846), specificity value (0.1667), precision value (0.6667) and F1-Score (0.4878). It can be deduced that the HESCA model with GBF outperforms that of the HESCA model with full-input features.

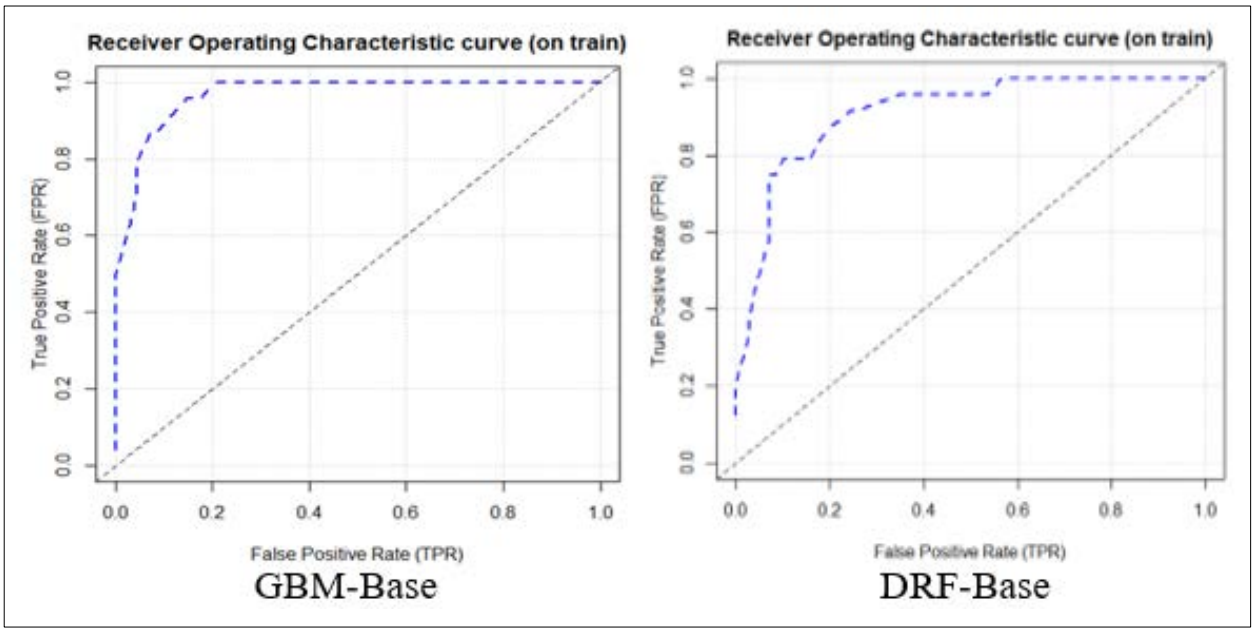
Table 8: Comparison of Base Models and HESCA Model Performance on Training Data based on GBF

	Base Models					Stacked Model
Metrics	**GBM**	**DRF**	**DNN**	**GLM**	**NB**	**HESCA Model**
Accuracy	0.9139	0.8279	0.8387	0.7957	0.7957	0.9677
Log loss	0.2838	0.5021	0.7200	0.4851	0.5926	0.1172
Recall	0.9000	0.7222	0.6552	0.6000	0.6000	0.9000
Specificity	0.9178	0.8533	0.9219	0.8677	0.8788	1.0000
Precision	0.7500	0.5417	0.7917	0.6250	0.6667	1.0000
F1-Score	0.8100	0.6191	0.7169	0.6122	0.6316	0.9474
AUC	0.9329	0.7416	0.8795	0.7769	0.7298	0.9952

Table 9: Comparison of Base Models and HESCA Model Performance on Test Data based on GBF

	Base Models					Stacked Model
Metrics	**GBM**	**DRF**	**DNN**	**GLM**	**NB**	**HESCA Model**
Accuracy	0.8438	0.7500	0.7188	0.7813	0.7500	0.9063
Log loss	0.4686	0.5156	0.7310	0.5038	0.4948	0.2959
Recall	0.6667	0.5385	0.5000	0.6667	0.5333	0.7500
Specificity	0.9500	0.8947	0.9375	0.8077	0.9412	1.0000
Precision	0.8889	0.7778	0.8889	0.4444	0.8889	1.0000
F1-Score	0.7619	0.6364	0.6400	0.5333	0.6667	0.8571
AUC	0.8285	0.7536	0.7778	0.8140	0.8019	0.9251

Table 8 and Table 9 show the comparative performance metrics of base models and the HESCA model on training and test data respectively based on GBF. It can be observed in Table 4 that the HESCA model with the least log loss value (0.1172) had the best accuracy (0.9677), specificity (1.00), precision (1.00), F1-Score (0.9474), and AUC (0.9952) compared to those of the base models. It is interesting to observe that the best recall value (0.9000) is recorded for both the GBM base model and the HESCA model. Figure 8 graphically shows the information in Table 8. It can also be observed in Table 9 that the HESCA model with the least log loss value (0.2959) had the best accuracy value (0.9063), recall (0.7500), specificity (1.00), precision (1.00), F1-Score (0.8571), and AUC (0.9251) on test set compared to those of base models. Figure 9 graphically shows the information in Table 9. In effect, it can be deduced that the HESCA model outperformed the base models on the train and test set based on the optimal feature subset of the data used in this study, indicating better predictions on patients with recurrent HNSCC prognosis. The ROC curve analyses of each base model and HESCA model on the training set and test set are shown in Figure 6 and 7 respectively.



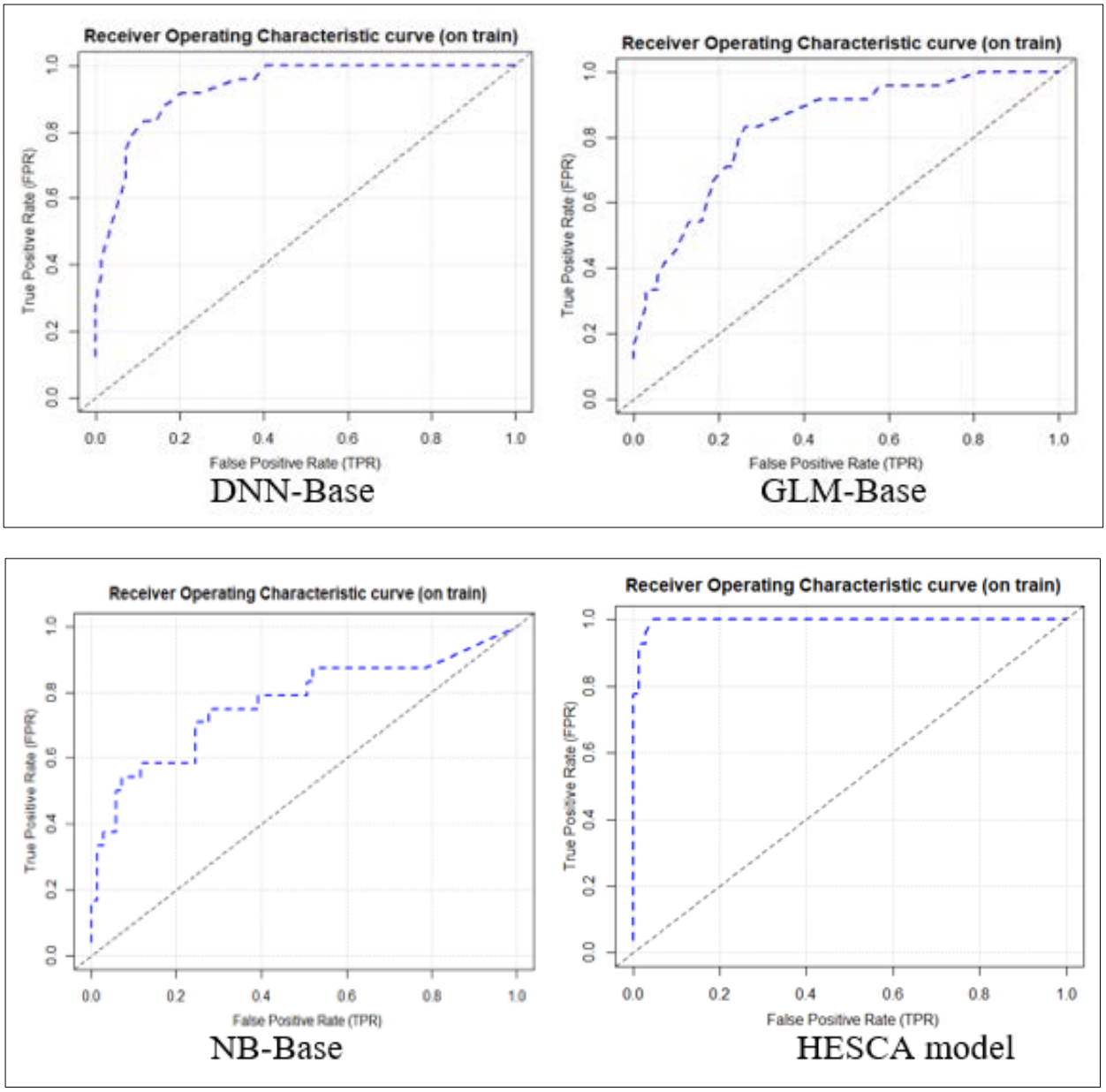
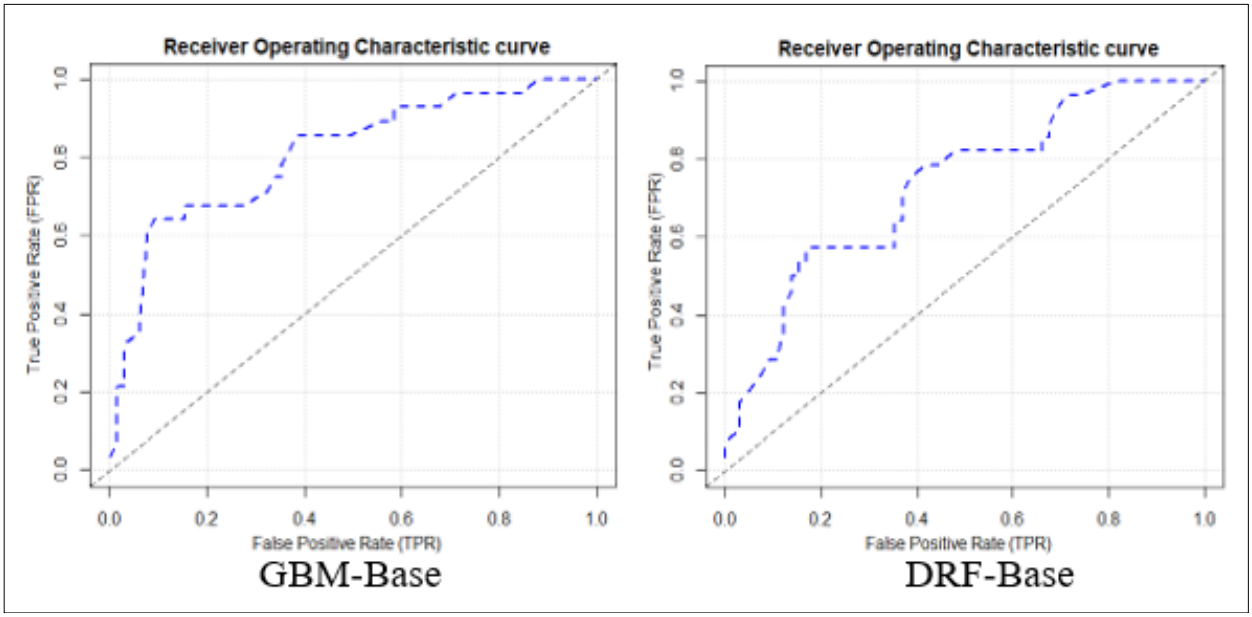


Fig 6: ROC Curve analysis of base models and HESCA model on the Training set



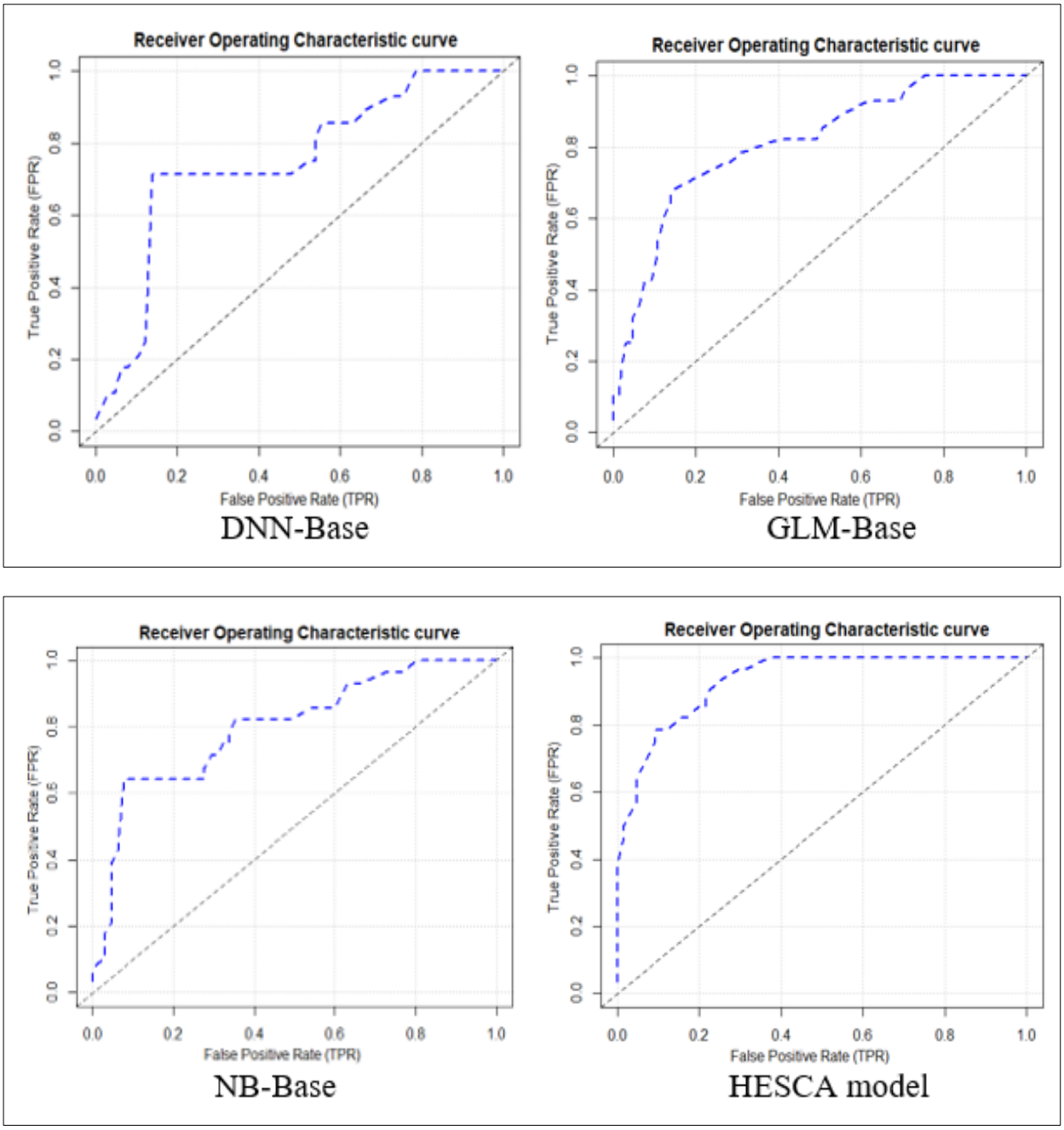


Fig 7: ROC curve analysis of base models and HESCA model on Test set

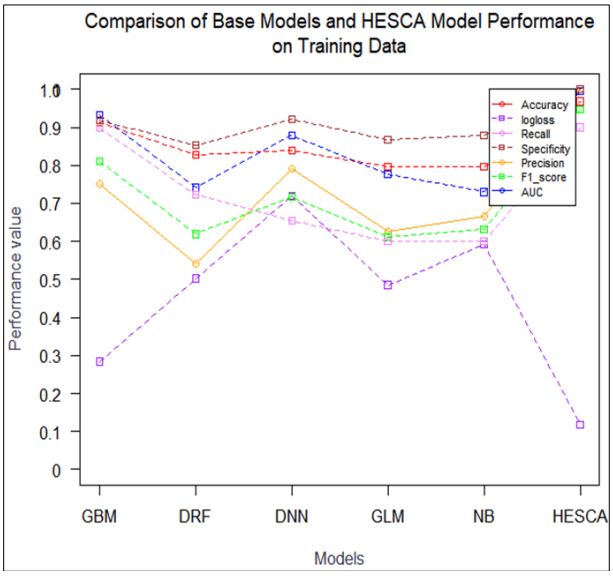


Fig 8: Graph of Base models and HESCA model on training set

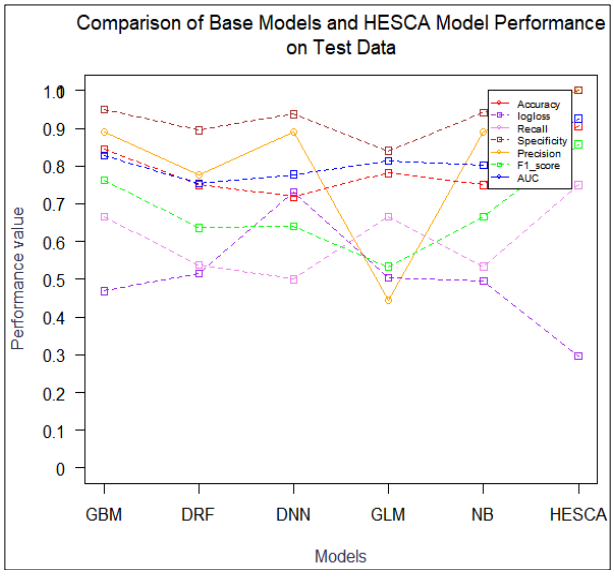


Fig 9: Graph of Based models and HESCA model on test set

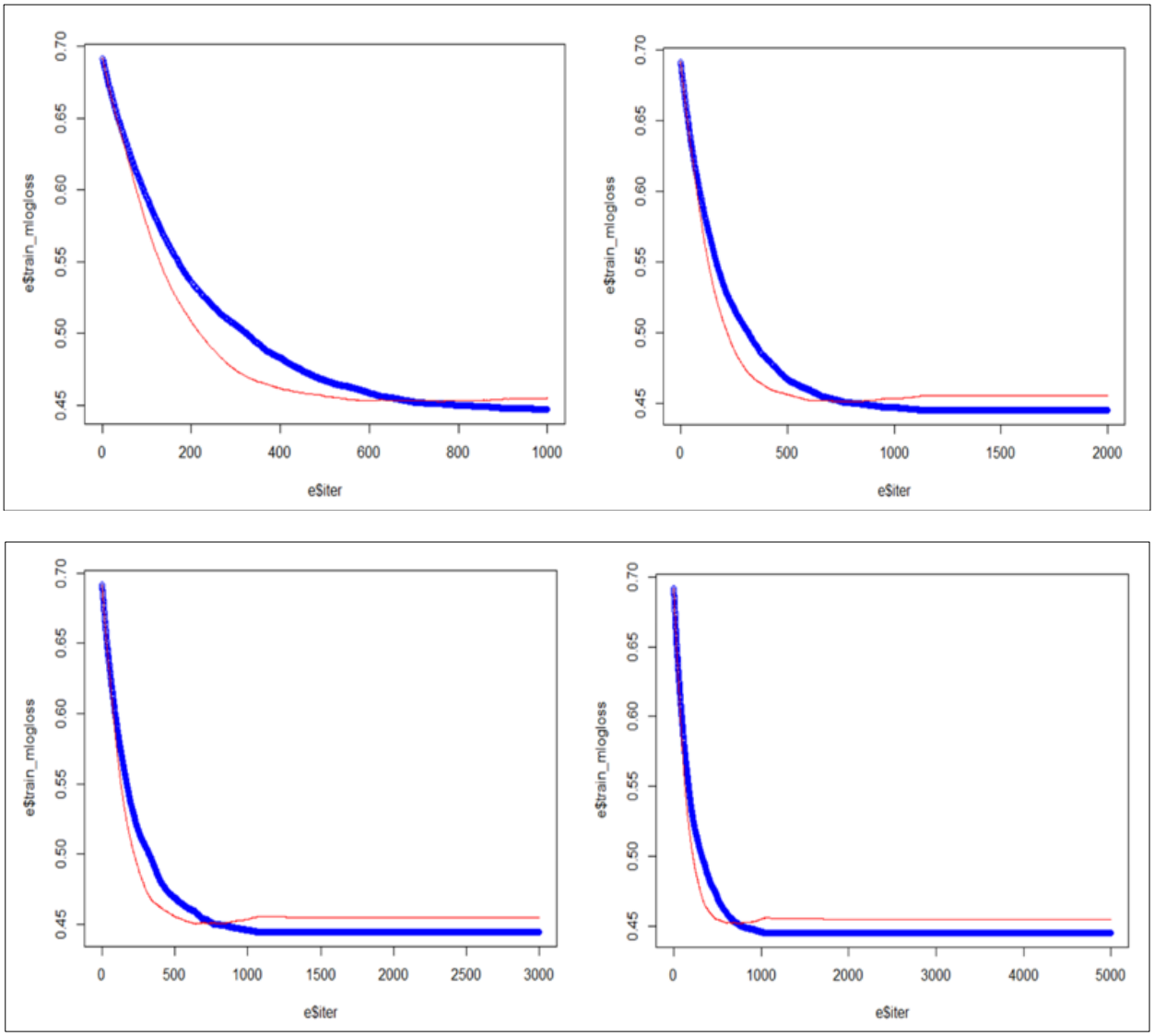


Fig 10: A good fit learning curves plot

Figure 10 displays the HESCA model's 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 *generalization 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 HESCA model suited the data.

Discussion

In Ghana, between the ages of 15 and 60, as opposed to after 60, HNSCC subtypes have a growing incidence rate [8]. This shows that the average age at which HNSCC develops is lower, and that the likelihood of patients being treated with the intention of curing their illness is very low because they may not be able to pay for the necessary care due to their poor financial situation. Where they are able to afford the curative intent treatment, they are not able to complete such treatment due to low-income level. Those that are able to complete such treatment for cancer to reach its remission, they still experience recurrence or relapse due to the advanced or metastatic stage of tumor at its diagnosis [8]. That, HNSCC recurrence is strongly linked to the stage of tumor at diagnosis [20]. In this case, early diagnosis and accurate prognosis is needed. A lot of research has recently been done in an effort to increase the precision of HNSCC prognosis and predictions using machine learning techniques, particularly the ensemble techniques that enhance the performance of the classification model by combining multiple different models rather than base models [4, 2]. In order to develop prediction models and have their performance tested in order to produce a more robust model, this study used the layered generalization (stacking) technique. Each of the strong base prediction models employed in this study-GBM, DRF, DNN, GLM, and NB-was used as a meta classifier in the stacking ensemble in order to create the best meta-model.

Even though the independent predictive models of GBM, DRF, DNN, GLM, and NB performed well on the GBF, performance improved across the board when the stacking strategy was used. The stacking ensemble utilizing GBM as a meta-classifier, on the other hand, demonstrated superior prediction accuracy than the pre-existing base models and other classifiers taken into consideration as meta-classifiers in this study. The outcomes of this study demonstrated that stacked ensemble models with GBM acting as a meta classifier are useful as a supplementary tool for categorizing and predicting recurrent HNSCC prognostic data. This gave the stacked ensemble model having five base models and a GBM meta-model based on GBF termed as HESCA model in this study.

Conclusion And Future Work

Stacked ensemble model termed as HESCA model through stacked generalization was presented for classifying and predicting recurrent HNSCC prognosis in random grid search and achieved significant improvement in its performance. GBFS provided 8-input features; *Nodes*, *Age*, *Smoke*, *StageIV*, *p63*, *TreatCCRT*, *PaTT4*, and *Size* as the optimal feature subset; thus, the most accurate prognosis for recurrent HNSCC. The proposed HESCA model (8-input model) based on GBF achieved significant improvement in accuracy by 56.25%, from 34.38% (model with full-input) to 90.63% with a reduction in log loss from 1.0435 to 0.2959. It was also observed that there is significant improvement in accuracy and log loss of GBM model from its base model to its stacked ensemble model with accuracy from 84.38% to 90.63% and log loss from 0.4686 to 0.2959.

In this work, the best meta classifier model was explored where the identical base models were utilized in both the base classifier and the meta classifier based on GBFS for the best feature subset. The findings of this work offer a viable guide for selecting machine learning models for additional stacking generalization research. It is anticipated that by incorporating many layers of stacking, it would be able to construct a two-layer multi-level stacked ensemble model with improved performance. However, this study has limitations because it is based on a one-layer stacked ensemble learning.

Data Availability

Contact the first author via email for data accessibility.

Acknowledgement

Nil

Disclosure statement

The authors reported no potential conflict of interest relevant to this study.

Funding

Self-funding

References

1. Abdulai EA, Nuamah IK. Incidence of squamous cell carcinoma of the oral cavity and oropharynx in Ghanaians - A Retrospective study of histopathological charts in a teaching hospital, World Journal of Surgical Medical and Radiation Oncology; c2013. p. 8.
2. Adeyemi OJ, Adebayo VO, Olaniyi O, Olusanya OO, Idowu PA. A Stack Ensemble Model for the Risk of Breast Cancer Recurrence, International Journal of Research Studies in Computer Science and Engineering. 2019;6(3):8-21.
3. Akinbohun F. A Stacked Ensemble Model for Diagnosis of Craniocervical (Head and Neck) Cancer, Post Graduate Research Unit of Albert Ilemobade Library, FUTA; c2021. p. 1.
4. Akinbohun F, Akinbohun A, Daniel A, Oyinloye OE. Diagnosis of Head and Neck Cancer in Developing Countries using a Stacked Ensemble Model, European Journal of Engineering Research and Science. 2020;5(9):1-5.
5. Alabi RO, Elmusrati M, Sawazaki-Colone I, Kowalski LP, Haglund C, Coletta RD, *et al*. Machine learning application for prediction of locoregional recurrences in early oral tongue cancer, A Web-based prognostic tool, VirchowsArchiv; c2019.
6. Bray F, *et al*. Global cancer statistics. Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians. 2018;6(68):394.
7. Chang S-W, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics. 2013;4:170.
8. Commeh E. Cancer Cases in Ghana are not decreasing, General News, Source: GNA; c2019.
9. Diyaol MH. Artificial Intelligence System to Predict and Diagnose Breast Cancer; c2018. diyaolihaqq@gmail.com
10. Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric decision support system for the prediction of oral cancer reoccurrence, IEEE Trans InfTechnol Biomed. 2011;16:1127-1134.
11. Gremmell D. Ensemble Learning in R with Super Learnerl; c2018. https://www.datacamp.com/tutorial/ensemble-r-machine-learning
12. Kabir F, Ludwig SA. Enhancing the Performance of Classification Using Super Learning, Data-Enabled Discov. Appl. 2019;3(5):2-13.
13. Kwon H, Park J, Lee Y. Stacking Ensemble Technique for Classifying Breast Cancer, Jnl of Health Informatics Research. 2019;25(4):283-288.
14. Kitcher ED, Yarney J, Gyasi RK, Cheyuo C. Laryngeal Cancer at the Korle-Bu Teaching Hospital Accra Ghana, Departments of Surgery and Pathology, University of Ghana Medical School, Department of Radiation Oncology, Ghana Medical Journal. 2006;2(40):1-5.
15. Lavanya L, Chandra J. Oral Cancer Analysis Using Machine Learning Techniques, International Journal of Engineering Research and Technology. 2019;12(5):596-601.
16. Larsen-Reindorf R, *et al*. A Six-Year Review of Head and Neck Cancers at the Komfo Anokye Teaching Hospital, Kumasi, Ghana, International Journal of Otolaryngology and Head & Neck Surgery. 2014;3:271-278.
17. Mitchell TM. The discipline of machine learning, Carnegie Mellon University, school of computer science, machine learning department; c2006.
18. Singh S, Yassine A, Benlamri R. Internet of Energy: Ensemble Learning through Multilevel Stacking for Load Forecasting, IEEE Intl Conf on Dependable; c2020. p. 664-668.
19. Wolpert DH. Stacked generalization, Neural Networks. 1992;5(2):241-259.
20. Worsham MJ. Identifying the risk factors for late-stage head and neck cancer: Expert review of anticancer therapy. 2011;11:1321-1325.
21. Yaliang L, Jing G, QI L, Wei F. Ensemble Learning; c2015. p. 484-504
22. Yarney, *et al*. Cancers of the Head & Neck: Does concurrent chemoradiotherapy preceded by chemotherapy improve survival in locally advanced nasopharyngeal cancer patients? Experience from Ghana, BioMed Central. 2017;2(4):1-7.