Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer

Abdelrahman Elsharif Karrar
College of Computer Science and Engineering
Taibah University, Medina, Saudi Arabia

***Abstract*—Class imbalance problem become greatest issue in data mining, imbalanced data appears in daily application, especially in the health care. This research aims at investigating the application of ensemble model by intelligence analysis to improving the classification accuracy of imbalanced data sets on prostate cancer. The primary requirements obtained for this study included the datasets, relevant tools for pre-processing to identify the missing values, models for attribute selection and cross validation, data resembling framework, and intelligent algorithms for base classification. Additionally, the ensemble model and meta-learning algorithms were acquired in preparation for performance evaluation by embedding feature selecting capabilities into the classification model. The experimental results led to the conclusion that the application of ensemble learning algorithm on resampled data sets provides highly accurate classification results on single classifier J48. The study further suggests that gain ratio and ranker techniques are highly effective for attribute selection in the analysis of prostate cancer data. The lowest error rate and optimal performance accuracy in the classification of imbalanced prostate cancer data is achieved using when Adaboost algorithm is combined with single classifier J48.**

***Keywords*—Ensemble model; intelligence analysis; classification of imbalanced data; prostate cancer**

I. INTRODUCTION

This Prostate cancer is among the leading causes of death in men worldwide. The prostate is a glandular structure located in the male productive system and its functions is to promote spermatic health and enhance fertility by adding a nutrient-rich alkaline fluid to the semen [1]. Malignant tumors that lead to prostate cancer state to develop when the rate of cell multiplication is higher than cell death. This alters the genetic structure leading to mutations and tumor metastasis on the urothelial lining. Compared to other glands, the prostate has a higher malignancy rate due to the heavy reliance on the androgenic signaling of hormones such as testosterone, abnormal Gli-1 oncogene expression, and Sonic Hedgehog (Shh) expression, which stimulate cellular proliferation and stromal tumor growth. The process in which prostate cancer develops is known as Prostatic Intraepithelial Neoplasia (PIN) While most research studies on the pathogenesis of prostate cancer report inconclusive findings, etiological factors such as genetic inheritance and family history, vasectomy, environmental carcinogens, low carotenoid intake, and high intake of saturated fats and other unhealthy dietary/lifestyle habits are known to increase the risks significantly.

Prostate cancer is classified as a carcinoma since its malignancy develops primarily from the epithelium lining of the peripheral glandular tissue. The epithelial structure of the prostate gland is composed of three cell types including rare neuroendocrine cells, basal cells, and luminal cells, which are responsible for the expression of androgen receptors, secretion of glycoprotein prostate specific antigen (PSA) and prostatic fluids [1]. Research studies suggest that prostate tumors that initially form from the luminal cells metastasize more rapidly compared to those from the basal cells due to the alteration of epithelial stromal tissues and the damage of glandular structure. The accuracy of clinical interventions such as the classification of diagnostic data from cancer tissues is influenced by a range of factors including the extent of cellular differentiation on histology and cyclic biochemical recurrence risk. Accurate classification of diagnostic data significantly influences the efficacy of treatment intervention through timely detection based on the Tumor, Nodes and Metastasis framework.

Data classification techniques for the diagnostic data are subject to structural imbalances and errors due to factors such as the underlying assumptions of evenly distributed training datasets. The classification approaches are highly vulnerable to bias when implemented on training data sets with severely imbalanced distribution. Insights from imbalanced training data sets may have severe practical implications on the associated decision outcomes. However, the problem of imbalanced data distribution is fairly common in real-world scenarios, especially when target classes lack uniform distribution across multiple class levels [2] . Data set imbalances occur when major classes have more instances and minor classes have relatively fewer instances. The classification of data sets with imbalanced distributions is a major challenge that has not been fully solved even by advanced machine learning algorithms with mathematical model mapping and computational prediction capabilities for identifying embedded data patterns [3].

This paper develops multiple potential approaches based on algorithmic modification, feature selection, ensemble learning, cost-sensitive learning, and sample selection methods to address the challenges of imbalanced distribution in learning data sets.

A. Production Statement

Class imbalance is the most occurring and potentially risky analytics issue, especially in the data mining of unstructured sets from healthcare systems and processes due to the high likelihood of some classes having larger sample sizes compared to others [4]. A significant number of the current data mining techniques are structurally designed to ignore misclassification risks on minor samples while focusing on the classification of major samples. The accuracy of data classification techniques is impeded by factors such as data imbalances coupled with uneven distribution and sample size differences from one class to another. As a result, traditional classification algorithms are highly unreliable and unsuitable due to high risks of bias and inaccuracy. This explains to need to determine and test whether a data classification model based on machine learning ensemble is capable of delivering comparatively higher levels of accuracy [5].

B. Research Questions

This research seeks to answer the following questions;

1) Can the implementation of machine learning ensemble model to data classification improve the classification of imbalanced data sets for prostate cancer management?

2) Is the application of resampling techniques based on machine learning reliable in optimizing and improving classification accuracy in imbalanced data sets for prostate cancer management?

This research paper is organized in sections including a review of recently published literature on classifiers and prostate cancer for comparisons with related studies in both fields in Section II, a detailed description of the experimental procedure, methodology, imputation process , and the general set up in Section III, and the evaluation of experimental results in Section IV. Finally, Section V of this research paper discusses conclusions based on the experimental results and provides recommendations for future studies.

II. Literature Review

This section provides a conceptual description of data mining with relation to techniques such as ensemble learning and resampling to investigate the implications of data classification accuracy on the management of prostate cancer, including a review of recently published literature on classifiers and prostate cancer for comparisons with related studies in both fields.

A. Prostate Cancer

According to 2021 prevalence statistics by the American Cancer Society, prostate cancer is the second most prevalence type cancer after skin cancer among men in the United States with approximately 248,530 new reported cases and about 34,130 deaths [6]. Data further shows that one in every 8 men develops prostate cancer, especially among adults aged above 65 of African ethnicity. Prostate cancer is ranked as having the seconding highest death rate from lung cancer in American males. Statistical estimates suggest that in a sample population of 41, one man dies of prostate cancer [6]. In addition to age, other risk factors for prostate cancer in men include family history through genetic inheritance, ethnicity (60% more risk among blacks), and lifestyle factors such as diet, smoking, and level of physical activity [6]. Early detection of prostate cancer is linked to significantly higher chances of survival and longevity. Studies suggest that timely detection of prostate cancer plays a significant role in the effectiveness of treatment interventions hence the need for various interventions to promote the identification and detection of early symptoms.

B. Data Mining Process

Data mining techniques are applied used to extract trends and patterns through the Knowledge Data Discovery process (KDD) [7]. The extraction of patterns among multiple variables depends on data mining techniques, which may be predictive or descriptive. Predictive data mining methods provide a generalized description of the data attributes while predictive data mining uses historical data to make accurate trend forecast [8].

Data Mining software are designed analyze data on the basis of parameters such as sequence analysis (a pattern in which events are interdependent), degree of association (where events defined by the datasets are interconnected), clustering (where data with identical patterns are grouped), and classification (where predefined variables are used to identify new patterns) [9].

C. Data Mining Techniques

The flow diagram shown in Fig. 1 provides a description of various techniques for data mining and retrieval based on regression, classification, clustering, and association [10].

D. Classification Techniques

The classification approach to data mining in healthcare entails predicting and grouping a data set in sample class categories [11]. This provides important insights for the identification of unique disease patterns that associate certain risk factors to a patient population through supervised learning [12]. Binary classification is a technique where the risk factors are classified as either ‘high’ or ‘low’ while multiclass technique involves more than two classes for example ‘high’, ‘medium’, or ‘low risks’ [8]. The data is further divided into classes; training and testing datasets, which are used to predict the possible outcomes from a historical event.

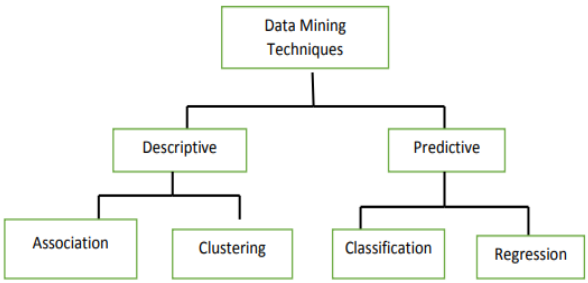

Fig. 1. A Flow Diagram Illustration of Data Mining Techniques.

1) Decision tree: Decision trees are used as classifier representations and are constructed using data to solve research problems such that attribute tests are denoted by non-leaf nodes and test outcomes denoted by branches while leaf nodes are assigned particular class levels [8]. Decision tree analysis is used by researchers to determine conditional probabilities for optimal decision making and class separation based on information gain. In the healthcare field, decision trees are used in the classification discrete values due to the ability to process nominal and numeric attributes while adjusting missing data values.

2) Support Vector Machine (SVM): Support vector machine is an advanced classification algorithm for linear and non-linear data sets. It is applied in the transformation of original training data to higher dimensions at which an optimal hyperplane that separates class instances can be determined. Support and marginal vectors provide a framework for determining the hyperplane in SVM subject to the kernel metric C = J [13].

3) Meta learning classifier: This is a classification approach in which historical data is used as a learning set using algorithms such as the random subspace, adaboost, and bagging. The adaboost algorithm is applied to improving the classification accuracy by performing multiple iterations to cluster weak learning algorithms and modifying the accuracy parameters, especially in imbalanced or misclassified sets. Adaboost algorithm is implemented as shown in Fig. 2 [14].

The bagging algorithm is implemented through boostrap aggregation, which involves deriving base classifiers from the decision tree. Boostrap samples D1, D2, …Dn are selected from a data set D provide the base classifiers C1, C2, … Cn [15]. Supposing that an optimal number of votes are assigned to a class for randomly selected labels, then the algorithm extracts training object and classifier sets for bootstrapping after which an integration process based on majority voting takes place [16]. The implementation procedure for the bagging algorithm is illustrated in Fig. 3.

Ensemble learning technique describes a process in which multiple classifiers are trained to generate decision insights based on different classifiers through random subspace, bagging, and boosting approaches for increased performance [17]. The most common ensemble learning approaches include weighted averages, majority voting, and simple averages. Ensemble techniques combines multiple classifiers in determining the optimal classification model from different sub-models comprising of a base classifier layer and meta-classifier layers, which make accurate predictions [17].

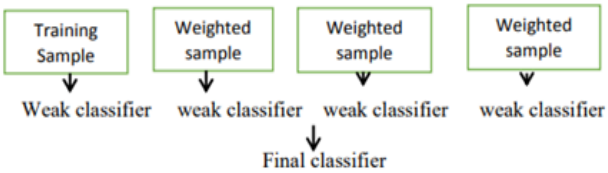

Fig. 2. The Implementation Stages of Adaboost Algorithm.

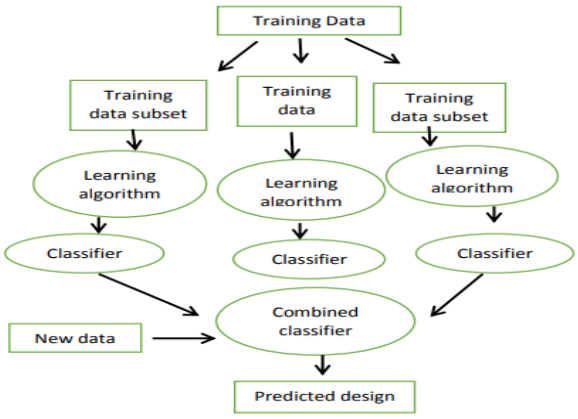

Fig. 3. The Implementation Stages of Bagging Algorithm Ensemble Learning [15].

4) Attribute subset selection: Attribute selection techniques play a significant role of data reduction for more efficient analysis in the data mining process. When data sets have many attributes, attribute selection is used to determine those that align to the cost of data analysis and utility for the easier discovery of patterns. Filter and wrapping categorization methods are used in evaluate the estimation accuracy of the learning algorithm [18].

5) Resampling, oversampling, and under sampling method: Data mining techniques are applied in healthcare to identify emerging trends from unstructured data sets. Resampling methods combine multiple approaches, which include the Random-oversampling of minor data classes, random oversampling of major classes hence providing solutions to sample distribution problems. Under-sampling removes data imbalances through the random elimination of major classes while oversampling achieves the same objective by replicating minor classes [19].

III. Methodology

This paper utilizes an integrated methodological framework for literature review, dataset extract, and pre-processing to prepare it for analysis. The primary requirements obtained for this project included the datasets, relevant tools for pre-processing to identify the missing values, models for attribute selection and cross validation, data resembling framework, and intelligent algorithms for base classification. Additionally, the ensemble model and meta-learning algorithms were acquired in preparation for performance evaluation by embedding feature selecting capabilities into the classification model.

A. Dataset Description and Data Transformation

The data used for this study was obtained from the prostate cancer unit at Mayo Clinic, Rochester from a sample population of 1144 patients whose attributes such as age, size of tumor, Node-caps, degree of malignancy, metastasis, and class were recorded. Data imbalances were detected in 808 zero reoccurrences and 336 recurrences as shown in the Table I.

TABLE I. DATASET DESCRIPTION

Attribute	Description	Attribute Type
Tumor	Swollen prostates	Numeric
Age	Age of the patient	Numeric
Node	Absence or presence of node	Nominal
Metastasis	Tumor spread throughout the body	Nominal
Class	Recurrence of risk factors	Nominal
Degree of malignancy	Stage of cancer development	Numeric

WEKA open source software was selected to perform the data mining processes in this study. This tool has integrated data mining capabilities for clustering, regression analysis, classification, pre-processing, and visualization [20]. Pre-processing was performed to ensure that the attribute types of each data class was either nominal or numeric and all missing values replaced with the computed average. Imbalance problems in the dataset were resolved through resampling techniques and the attributes were selected through a dimensionality reduction technique with optimal gain ratio.

B. Classification Algorithms Selection

The classifier algorithms selected for data mining in this study include J48, Neural Networks, Rep Trees, and SVM as the base classifiers, and meta-classifiers such as random subspace, boosting, and bagging, which were used in the building of classifier models.

1) Decisions tree J48 algorithm: The basic algorithm involved processes such as the construction of decision trees using the top-down divide-and-conquer approach with root training examples based on the categorical classification of attributes [21]. Recursive partitioning of the heuristic measures was implemented under conditions that all samples are assigned to the same classes and leaf classification based majority voting for all samples [22].

2) Neural network algorithm: The data input is embedded simultaneously into input layer after which it is weighted and adopted to a hidden layer, which is usually arbitrary. The last hidden layer contains weighted outputs which form the output layer, which produce predictive insights about the network patterns [20]. A feed-forward approach is applied to the network such that weight cycles in the input or output units are not returned to their previous layer.

3) Rep tree algorithm: This algorithm prunes the decision tree to allow the re-generation of the initial tree with minimal error. The data instances are segmented into multiple units such that set leaf can be assigned the lowest number of instances [23].

4) SVM algorithm: A relatively new classification method for both linear and nonlinear data, It uses a nonlinear mapping to transform the original training data into a higher dimension [24].

With the new dimension, it searches for the linear optimal separating hyperplane (i.e. “decision boundary”).

With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.

SVM finds this hyperplane using support vectors (“essential” training tuples) and margins (defined by the support vectors).

5) Bagging algorithm: This algorithm classifies datasets into training and testing categories. Multiple training sets are generated and replaced in iterative sequences to reduce the likelihood of over-fitting and control variance.

6) Boosting algorithm: The implementation of this algorithm follows an iterative procedure for adaptive classification of training datasets, especially the misclassified sets. The algorithm assigns equal weights to the initial records N and performs automatic adjustments unit weights are increased in the wrongly classified datasets and increased in the accurately classified data sets [25].

7) Random subspace algorithm

Repeat forb= 1, 2, . . . ,B.

Choose an r-dimensional random subspace b from the original p-dimensional feature space X.

Build a classifier Cb(x) (with a decision boundary Cb(x) 0) in b.

Aggregate classifiers Cb(x),b=1, 2, . . . ,B, by majority voting for the final decision. [26].

C. Ensemble Learning

Ensemble model was applied to a combination of classifiers to determine the point at which classification performance is optimal. Ensemble learning model is composed of the base classifier and meta-classifier layers which receive and analyze prediction inputs to generate the desired output.

D. Evaluation Approach and Techniques

The ensemble model is utilized to classify prostate cancer data using combined sub-classifiers to improve performance and accuracy. Factors such as the relative accuracy of measures, degree of training and simulation errors, and classifier performance are used to validate the model [27]. Recall and precision measures are used to determine the accuracy of classification techniques [28]. Additionally, each classifier is evaluated on the basis of computation time matrix, which shows the rate at which algorithms make correct and incorrect predictions compared to the actual values defined in the dataset [29]. The evaluation of metrics accuracy is illustrated in the Table II.

TABLE II. EVALUATION OF CONFUSION METRICS ACCURACY

	Positive Prediction Class	Negative Prediction Class
Real Class Positive	True Positive	False Negative
Real Class Negative	False Positive	True Negative

True Positive: Accurate classification of recurrence instances.

True negative: Inaccurate classification of no-recurrence instances.

False positive: Inaccurate classification of no recurrence instances as recurrent instances.

False negative: Inaccurate classification of recurrence instances no-recurrence instances.

In order to get TP rate, FP rate, Precision, Recall, F-Measure, Accuracy were used in this research as follows:

1) True Positive (TP) rates (sensitivity/recall) – is the proportion of the actual recurrence (or no recurrence) cases correctly classified.

TP (recurrence) = *TP*/ (TP +FN)

TP (no recurrence) = TN / (TN +FP)

2) False Positive (FP) rates (1-specificity/false alarms) – proportion of actual no recurrence (or recurrence) cases misclassified.

FP (recurrence) = FP/ (FP + TN)

FP (no recurrence) = FN/ (FN + TP)

3) Precision – proportion of predicted recurrence (or no recurrence) cases that were correct classified.

Precision (recurrence) = TP / (TP + FP)

Precision (no recurrence) = TN / (FN + TN)

4) Recall–Proportion of predicted recurrence (or no recurrence) cases that were correct classified.

Recall = TP / (TP+FN)

5) F—one of the performance measures that is used to retrieve data:

F-measure = 2 x Precision x Recall / Precision + Recall = 2 x TP / (2 x TP + FP + FN)

6) Accuracy – proportion of the total predictions that was correct.

Accuracy = TP + TN / (TP + TN +FP +FN). [30]

E. Receiver Operation Characteristic (ROC) Curve

ROC curves are used in the summarization of classifier performance based on the analysis of error rates involving false positives and true positives. Acceptable performance metrics are defined by the area under curve and represents the optimal decision boundaries for measuring the estimated costs of instance misclassification [31].

IV. Experimental Procedures and Results

This section discusses the experimental procedures involving base classifiers with selected attributes, without selected attributes and resampling method in the first experiment while the second involved meta classifiers with selected attributes, without selected attributes and resampling method.

A. Dataset

The dataset used for these experiments was obtained from the prostate cancer department at Mayo Clinic. The instances are defined by the attributes defined in the methodology section. The data was pre-processed and analyzed using WEKA software. The table shown in Fig. 4 shows how the dataset appeared after preparation using the software.

A	B	C	D	E	F	G	H	I	J
age	tumor-siz	node-caps	deg-malig	breast	Metastasi	irradiat	Class		
29	3.6	yes	3	R	yes	no	recurrence-events		
49	3	no	1	R	no	no	no-recurrence-events		
53	4.8	no	2	L	yes	no	recurrence-events		
26	1.6	yes	3	R	yes	yes	no-recurrence-events		
40	1.8	yes	2	L	yes	no	recurrence-events		
36	0.8	no	2	R	no	yes	no-recurrence-events		
34	4.2	yes	3	L	no	no	no-recurrence-events		
36	1.9	no	2	L	yes	no	no-recurrence-events		
26	4.8	no	2	R	no	no	no-recurrence-events		
26	1	yes	2	R	yes	yes	no-recurrence-events		
49	1.1	no	2	L	no	no	no-recurrence-events		
52	4.8	no	2	R	yes	no	no-recurrence-events		
59	2.1	no	1	R	yes	no	no-recurrence-events		
26	2.5	yes	2	R	no	no	no-recurrence-events		
53	2.7	no	2	L	yes	yes	recurrence-events		
57	1.4	no	3	L	no	no	no-recurrence-events		
60	1.9	yes	1	R	no	no	no-recurrence-events		
49	2.9	yes	2	R	yes	no	no-recurrence-events		
37	3.5	no	2	L	no	no	no-recurrence-events		
26	2.7	no	3	L	no	no	no-recurrence-events		
34	2.1	no	1	L	no	no	recurrence-events		
26	2.3	no	2	R	no	yes	no-recurrence-events		

Fig. 4. Sample of Data After Preparing.

B. First Experiments and Results

The first experiment involving base classifiers was conducted to investigate the performance of different algorithms involving imbalanced prostate cancer data sets. The algorithms were applied to data through sampling techniques, without attribute selection, and with attribute selection.

1) Result of support vector machine: The implementation of SVM algorithm to data classification without attribute selection had a performance accuracy of 70.63% within duration of 0.15 seconds, 0.3 mean absolute error, kappa statistic 0, relative absolute error 70.8%, and 118.99% root relative squared error as shown in the Fig. 5.

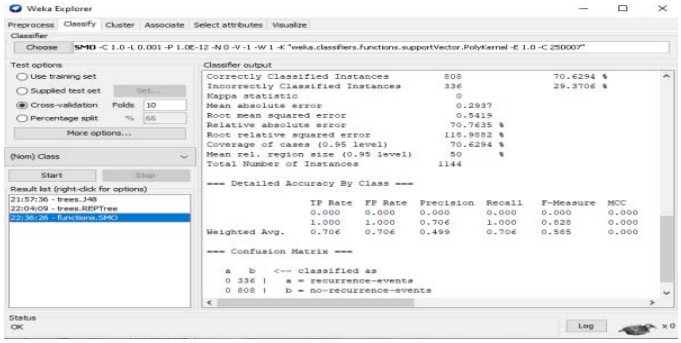

Fig. 5. Result of Classification Model using SVM without Attribute Selection.

The performance outcomes of other base classifier algorithms are shown in the Table III.

TABLE III. RESULTS OF A BASE CLASSIFIER ALGORITHMS WITHOUT ATTRIBUTE SELECTION

Evaluation Criteria	Classifier			
	J48	*SVM*	*ANN*	*Rep Tree*
Duration (seconds)	0.12	0.15	1.54	0.06
Correct classification	799	808	771	792
Incorrect classification	345	336	373	352
Percentage Accuracy	69.8%	70.63%	67.39%	67.39%

The experimental results of algorithm implementation of base classifiers with attribute selection are shown in the Table IV.

TABLE IV. RESULTS OF A BASE CLASSIFIER ALGORITHMS WITH ATTRIBUTE SELECTION

Evaluation Criteria	Classifier (With ranker and gain ratio)			
	J48	*SVM*	*ANN*	*Rep Tree*
Duration (seconds)	0	0.14	1.24	0.06
Correct classification	799	808	771	792
Incorrect classification	345	336	373	352
Percentage Accuracy	69.8%	70.63%	67.39%	69.23%

The performance of experimental parameters was evaluated based on criteria such as the mean errors and kappa statistic is shown in the Table V.

TABLE V. SIMULATION RESULTS

Evaluation Criteria	Classifier (With Ranker and Gain Ratio)			
	J48	*SVM*	*ANN*	*Rep Tree*
Kappa Statistic	0.0038	0	0.035	0.024
Mean Absolute error	0.414	0.294	0.411	0.042
Root mean squared error	0.462	0.542	0.472	0.472
Relative absolute squared error	99.78%	70.76%	99.06%	100.39%
Root relative squared error	101.42%	118.98%	103.63%	103.67%

2) Experiment using rep tree with resampling method: Resampling technique was applied on the base classifiers and implemented on decision tree rep to obtain accuracy scores in the data classification. The implementation results for each classifier algorithm are shown the Table VI.

C. Second Experiment and Results

The second experiment involved the analysis of performance classification scores on meta learning algorithms with and without attribute selection. The relative accuracy values are shown in the Table VII and Table VIII.

1) Evaluation of algorithms: The algorithms were further evaluated using criteria such as the mean errors and Kappa statistics and the results are shown in the Table IX.

TABLE VI. REP TREE WITH RESAMPLING METHOD

Evaluation Criteria	Classifier (With Resampling)			
	J48	*SVM*	*ANN*	*Rep Tree*
Duration (seconds)	0	0.06	0.93	0.01
Correct classification	799	808	771	792
Incorrect classification	345	336	373	352
Percentage Accuracy	69.84%	70.63%	71.24%	77.27%

TABLE VII. RESULTS OF META CLASSIFIERS WITHOUT ATTRIBUTE SELECTION

Evaluation Criteria	Classifier		
	Bagging	*Boosting*	*Random Subspace*
Duration (seconds)	0.58	0.47	0.2
Correct classification	795	808	807
Incorrect classification	349	336	337
Percentage Accuracy	64.9%	70.63%	70.54%

TABLE VIII. RESULTS OF META CLASSIFIERS WITH ATTRIBUTE SELECTION

Evaluation Criteria	Classifier		
	Bagging	*Boosting*	*Random Subspace*
Duration (seconds)	0.27	0.17	0.13
Correct classification	795	808	807
Incorrect classification	349	336	337
Percentage Accuracy	64.9%	70.63%	70.54%

TABLE IX. SIMULATION RESULTS

Evaluation Criteria	Classifier		
	Bagging	*Boosting*	*Random Subspace*
Kappa Statistic	0.045	0	-0.001
Mean Absolute Error	0.414	0.411	0.413
Root mean squared error	0.4665	0.4555	0.4553
Relative Absolute error	99.7%	98.95%	99.54%
Root relative squared error	102.42%	100.01%	99.96%

V. RESULTS AND DISCUSSION

The experimental results suggest that a combination of boosting and bagging classification algorithms achieve a higher level of accuracy when resampling is applied to SVM and rep tree. Resampling method effectively improve the accuracy of ensemble learning model when applied to imbalanced datasets on prostate cancer [32]. When each the performance of each algorithm is analyzed after resampling, algorithms with single classifiers such as SVM, neural network, rep, and J48 are more accurate and require less computational time. The experimental outcomes of all trials suggest that the implementation of ensemble learning model yields higher classification accuracy

on J48 tree after resampling compared to before resampling imbalanced datasets. The relative performance of each base classifier on the ensemble model under different conditions of resampling is shown in the Fig. 6.

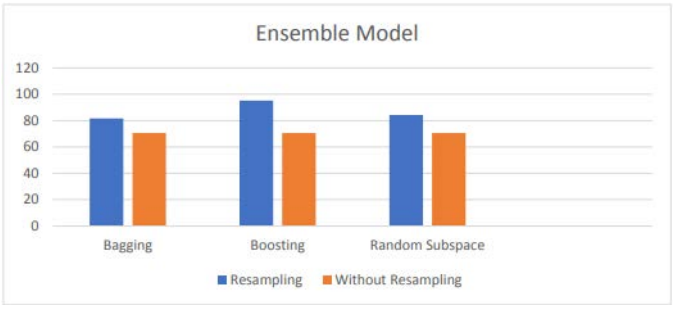

Fig. 6. Comparative Graph for different base Classifiers with different Evaluation Accuracy of Ensemble Model.

VI. Conclusion

The objective of this study was to develop a classification model for imbalanced prostate cancer datasets from Mayo Clinic, Rochester. The implementation of accurate classification approaches is important in the early detection and prediction of likelihood of recurrence or no-recurrence of risk factors. The experimental results led to the conclusion that the application of ensemble learning algorithm on resampled data sets provides highly accurate classification results on single classifier J48. The study further suggests that gain ratio and ranker techniques are highly effective for attribute selection in the analysis of prostate cancer data. The lowest error rate and optimal performance accuracy in the classification of imbalanced prostate cancer data is achieved using when Adaboost algorithm is combined with single classifier J48.

The following recommendations were developed based on the empirical results obtained from this study; Consider larger datasets to improve the accuracy of results, implement multiple evaluation techniques, and formulate alternative prediction models and algorithms to allow for the comparative analysis of classification results for imbalanced data.

References

[1] B. Murray, "The Pathogenesis of Prostate Cancer," in Prostate Cancer, Xon Publications, 2021, pp. 29-41.

[2] S. Saeed and H. C. Ong, "A bi-objective hybrid algorithm for the classification of imbalanced noisy and borderline data sets," Pattern Analysis and Applications, vol. 22, pp. 979-998, 2019.

[3] A. Elhassan, M. Aljourf, F. Al-Mohanna and M. Shoukri, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with random under-sampling (RUS) as a data reduction method," Global Journal of Technology & Optimization, 2016.

[4] A. S. Pranto and M. K. Paul, "Performance Analysis of Ensemble Based Approaches to Mitigate Class Imbalance Problem after Applying Normalization," in 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021.

[5] M. Ye and C. Wu, "Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble," Computational and Mathematical Methods in Medicine, vol. 2021, pp. 1-12, 2021.

[6] C. H. Pernar, E. M. Ebot, K. M. Wilson and L. A. Mucci, "The Epidemiology of Prostate Cancer," Cold Spring Harb Perspect Med, vol. 8, no. 12, 2018.

[7] A. E. Karrar, "The Use of Case-based Reasoning in a Knowledge-based (Learning) Software Development Organizations," International Journal of Innovative Research in Science, Engineering and Technology, vol. 5, no. 5, 2016.

[8] P. Ahmad, S. Qamar and S. Q. A. Rizvi, "Techniques of Data Mining In Healthcare: A Review," nternational Journal of Computer Applications, vol. 15, pp. 38-50, 2015.

[9] N. Padhy, P. Mishra and R. Panigrahi, "The Survey of Data Mining Applications And Feature Scope," International Journal of Computer Science, Engineering and Information Technology, vol. 2, no. 3, pp. 43-58, 2012.

[10] R. H. Alsagheer, A. F. Alharan and A. S. A. Al-Haboobi, "Popular Decision Tree Algorithms of Data Mining Techniques: A Review," International Journal of Computer Science and Mobile Computing, vol. 6, no. 6, pp. 133-142, 2017.

[11] M. Mutasim and A. Karrar, "Impute Missing Values in R Language using IBK Classification Algorithm," International Journal of Engineering Science and Computing, vol. 11, no. 6, pp. 28328-28338, 2021.

[12] A. E. Karrar, "A Novel Approach for Semi Supervised Clustering Algorithm," International Journal of Advanced Trends in Computer Science and Engineering, vol. 6, no. 1, pp. 1-7, 2017.

[13] M. A. Yaman, A. Subasi and F. Rattay, "Comparison of Random Subspace and Voting Ensemble Machine Learning Methods for Face Recognition," Symmetry, vol. 10, no. 11, 2018.

[14] L. Zhao, Z. Shang, A. Qin, T. Zhang, L. Zhao, Y. Wei and Y. Y. Tangd, "A cost-sensitive meta-learning classifier: SPFCNN-Miner," Future Generation Computer Systems, vol. 100, pp. 1031-1043, 2019.

[15] N. Joshi and S. Srivastava, "Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees)," International Journal of Computer Science and Mobile Computing, vol. 3, no. 5, pp. 727-732, 2014.

[16] D. P. Rangasamy, S. Rajappan, A. Natarajan, R. Ramasamy and D. Vijayakumar, "Variable population-sized particle swarm optimization for highly imbalanced dataset classification," Computational Intelligence, vol. 37, pp. 873-890, 2021.

[17] M. Mohammed, H. Mwambi, B. Omolo and M. K. Elbashir, "Using stacking ensemble for microarray-based cancer classification," in International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018.

[18] T.-B. A.J., C. L. and L.-D. R., "Attribute Subset Selection for Image Recognition. Random Forest Under Assessment," in 16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021). SOCO 2021. Advances in Intelligent Systems and Computing,, 2022.

[19] H. Li and M. Zhuang, "Clustering Center Optimization under-Sampling Method for Unbalanced Data," Journal of Software, vol. 15, no. 3, pp. 74-85, 2020.

[20] J. Nuhić and J. Kevrić, "Prostate Cancer Detection Using Different Classification Techniques," in CMBEBIH 2019, IFMBE Proceedings, Springer, Cham., 2020.

[21] A. M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," in 2020 3rd International Conference on Intelligent Sustainable Systems , 2020.

[22] M. F. Maulana and M. Defriani, "Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period," Journal Penelitian Ilmu Komputer, System Embedded & Logic, vol. 8, no. 1, pp. 39-48, 2020.

[23] R. Naseem, B. Khan, A. Ahmad, A. Almogren, S. Jabeen, B. Hayat and M. A. Shah, "Investigating Tree Family Machine Learning Techniques for a Predictive System to Unveil Software Defects," Complexity, vol. 2020, no. 6, pp. 1-21, 2020.

[24] N. Hafidz, Sfenrianto, Y. Pribadi, E. Fitri and Ratino, "ANN and SVM algorithm in Divorce Predictor," International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 2523-2527, 2020.

[25] M. M. Nishat, T. Hasan, S. M. Nasrullah, F. Faisal, A.-A.-R. Asif and A. Hoque, "Detection of Parkinson's Disease by Employing Boosting Algorithms," in 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 2021.

[26] M. A. Corsetti and T. M. Love, "Grafted and vanishing random subspaces," in Pattern Analysis and Applications, Springer, 2021.

[27] A. Doganer, "Different Approaches to Reducing Bias in Classification of Medical Data by Ensemble Learning Methods," International Journal of Big Data and Analytics in Healthcare, vol. 6, no. 2, pp. 15-30, 2021.

[28] M. Umair, F. Majeed, M. Shoaib, M. Q. Saleem, M. S. Adrees, A. E. Karrar, S. Khurram, M. Shafiq and J.-G. Choi, "Main Path Analysis to Filter Unbiased Literature," Intelligent Automation and Soft Computing, vol. 32, no. 2, pp. 1179-1194, 2022.

[29] C. N.V., Data Mining for Imbalanced Datasets: An Overview., Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA., 2009, pp. 875-886.

[30] P. Cristaldo, D. D. Luise, L. L. Pietra, A. D. Battista and D. Hemanth, "Data Mining-Based Metrics for the Systematic Evaluation of Software Project Management Methodologies," EAI/Springer Innovations in Communication and Computing. Springer, Cham., 2022.

[31] G. B. Demisse, T. Tadesse and Y. Bayissa, "Data Mining Attribute Selection Approach for Drought Modelling: a Case Study for Greater Horn Of Africa," International Journal of Data Mining & Knowledge Management Process, vol. 7, no. 4, pp. 1-16, 2017.

[32] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin and Y. Jin, "Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection," Applied Soft Computing, vol. 77, pp. 1-37, 2019.