2022 25th International Conference on Computer and Information Technology (ICCIT)
17-19 December, Cox's Bazar, Bangladesh

Machine Learning Approaches to Metastasis Bladder and Secondary Pulmonary Cancer Classification Using Gene Expression Data

Shovito Barua Soumma[1*], Ishraq R. Rahman[1*], Faisal Bin Ashraf[2]
[1]Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Bangladesh
[2]Department of Computer Science and Engineering, Brac University, Bangladesh
Email: [1]1605066@ugrad.cse.buet.ac.bd, [1]ishraq.r.rahman@gmail.com, [2]faisal.ashraf@bracu.ac.bd

***Abstract*—Similar causal relationships can exist between many cancer types, for example, metastatic bladder cancer and secondary lung cancer. This relatedness must therefore be taken into account for the diagnosis to be more accurate. The categorization of cancers can benefit from gene expression studies. In order to categorize cancer tissues with a comparable causal link, the best classifier model is sought after in this research. The CuMiDa dataset is used to obtain the lung and bladder cancer datasets, and parameters are modified to improve accuracy once fewer classifiers are taken into account. According to the experimental findings, Linear SVC achieves the highest accuracy, followed by Logistic Regress and XGBoost.**

***Index Terms*—Bladder Cancer, Lung Cancer, CNN, Deep Neural Network, Machine Learning, SVM, Feature Extraction**

I. Introduction

Classification of cancerous tumors poses great importance in the medical field. Working with an accurate prediction of different tumor types will certainly go a long way in better treatments as well as reducing misdiagnosis. But cancer classification is not that simple. Generally, cancer survivors might be at an increased risk of developing second primary malignancies. This is in particular due to the fact that cancers might share a common risk factor like smoking [1] [2]. Hence, we have to take into consideration the links that might exist in different forms of cancer types if we are to work on increasing accurate detection.

Moreover, cancer classification can be greatly benefited by adopting a systematic approach based on global gene expression analysis [3]. Thousands of genes can now be simultaneously monitored due to the advancements in microarray technology. Previously, a lot of cancer classification models had been proposed but in recent years researchers have been looking into cancer classification using gene expression as it has been shown that gene expression changes are related to different types of cancers [4]. This approach stands out from the rest because of its unique nature and domain of application.

Therefore, the objective of this work is to provide a classification system for cancer kinds that may share comparable causal relationships. The forms of lung and bladder cancer chosen for this article share the same causal relationship, which has been established. [5]. In this paper our contributions are -

- We considered cancer classification based on gene expression differences between cancer types with similar causal links.
- We conducted experiments to determine the best classifier model for accurate detection.

The paper is organized in the following way: In section II, an array of related works is presented to create an understanding of the background of our problems. In section III, the problem definition is established, and then the approach to resolving the problem is stated. Section IV goes on to draw a comparative analysis between the obtained result from classifiers and come to a verdict based on accuracy. Finally, section V states the future work possibilities and draws a conclusion to the paper.

II. Related Works

The process of transcribing a gene's DNA sequence into RNA is called *gene expression*. Certain diseases like cancer can be reflected in the change of expression values in certain genes. Hence, in [3], DNA micro-array expression data is used to classify a cancer type. Several classifier models were run on such data to gain an understanding of which model works the best for gene expression. Here, both classification accuracy and biological relevance were given the same amount of priority. In the end, it was found that for leukemia and ovarian data sets SVM has the highest accuracy, whereas, CAST performs better on the colon dataset. On the other hand, for biological relevance, no model is found to be superior to the others and hence, more exploration is left intended for future works.

In another work [6], it is shown that genetic expression plays a vital role in determining both phenotype and clinical course. Using this idea as a basis, bladder tumors are divided into more homogeneous and clinically relevant subgroups. Hence, genes that make the bladder tumors distinct will be majorly important to find the development and progression of bladder cancer. Kuper et al [1] try to establish a causal association between tobacco use and the risk of specific cancer types. Multiple different forms of cancer types will bear an increased amount of risk due to tobacco use. Among them, the lung, laryngeal, bladder, etc are included.

[*]***These authors contributed equally to this work***

979-8-3503-4602-2/22/$31.00

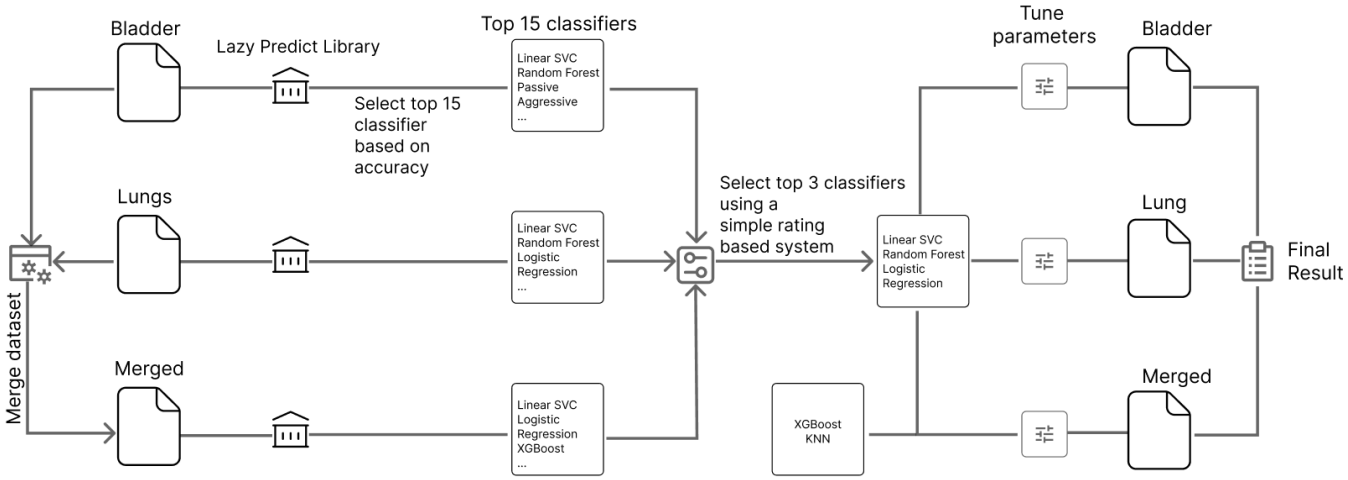


Fig. 1: Workflow of the proposed method for finding a better classifier

Aggressive bladder cancer treatments sometimes face obstacles since there lies a possibility of discovering indeterminate lung lesions on staging imaging. These findings may represent the presence of metastatic bladder cancer or a second primary lung cancer. Taylor et al [5] demonstrate that through a population-based assessment, all bladder cancer patients face an increased risk of synchronous and metachronous lung cancers. On the other hand, adults with muscle-invasive bladder cancer's lung nodules will more likely be metastatic bladder cancer rather than a second lung primary.

Abdullah et al [7] provides an effective method to predict lung cancer in an early stage with high accuracy ratio. In this paper, three classifier models (SVM, CNN & KNN) are run to inspect the accuracy ratio among them by using the WEKA tool. SVM achieves the highest accuracy whereas KNN has the least. The association between gene expression and the development of various types of cancer has been discovered utilizing a variety of data mining approaches. In a recent study, the authors [8] looked at various distance metrics to determine the best way to group genes based on their expression data in order to study how the genes affect various tumors.

III. Methodology

A. Dataset

In this paper, the Curated Microarray Database (CuMiDa) [9] is the source for the required data. Here, GSE 19804 dataset is going to be used for lung cancer and GSE 31189 for bladder. For GSE 19804, RNA was extracted from paired tumors and normal tissue for gene expression analysis. The report is a comprehensive analysis of the molecular signature of non-smoking female lung cancer in taiwan. On the other hand, for GSE 31189, total RNA was extracted from 52 samples of urothelial cancer cells and from 40 samples of non-cancer cells. Exfoliated urothelia from patients with known bladder disease status were used to apply differential gene expression analysis, then using a quantitative PCR approach selected targets from the microarray data were validated in an independent set of samples. Details for each of these datasets are shown in Table I:

TABLE I: List of GSEs and dataset employed in this work.

Datasets	Cancer Type	Samples	Genes	Classes
GSE31189	Bladder	85	54676	2
GSE19804	Lung	114	54676	2

B. Data Preprocessing

Since, the goal is to find the approach that best classifies cancer types that might have similar causal links, 3 types of datasets need to be considered as shown in Fig. 1.

- Individual Dataset for Bladder **(ID-B)**
- Individual Dataset for Lung **(ID-L)**
- Merged Dataset **(MD)**

Initially, the bladder and lung cancer datasets were collected from CuMiDa. These are the "Individual Dataset". After that, since we need to consider the correlation between lung cancer and bladder cancer from the perspective another case need to be considered where both cancer tissues are considered with non-tumoral tissue. Through this approach, "Merged Dataset" is achieved.

C. Proposed Approach

Our goal is to find the best suited classifiers to distinguish between bladder and lung cancer tissue using the newly created datasets mentioned in III-B. But there are several classifiers that might be perfect for the job. Hence, before landing on a set of classifiers for parameter tuning to find the best possible result, it is essential to entertain the myriad of options out there. That is where, the *"Lazy Predict Library"* comes into play.

1) Lazy Predict Library: It is often really hard to select the best ML models suited for a task as it requires a lot of time to go through each model, tune them to a certain specification and achieve higher accuracy. "Lazy Predict Library" intends

to overcome this hurdle. This module helps us find the best model for classification in a lazy manner. Even though it might require high computational power and might be a little time-consuming for high-dimensional data with multiple features, it is still a better approach to narrow down the best classifier options over the brute approach.

2) Classifier Selection: Initially, the lazy predict library is run for the three datasets (*ID-B, ID-L, MD)*. From there, the top 15 classifiers for each dataset were chosen, and then using a simple rating-based system based on the accuracy score for each classifier and dataset, three classifiers were chosen. These are : *Linear SVC, Logistic Regression & Random Forest* . Additionally, KNN and XGBoost classifiers are also going to be considered for *MD* so as to give credence to the results obtained from Lazy Predict Library are indeed correct. The results obtained from the library are listed in Table II.

TABLE II: Results obtained from "Lazy Predict Library"

Individual Dataset For Bladder (*ID-B*)

Model	Accuracy	Balanced Accuracy	F1 Score	Time Taken
Linear SVC	0.77	0.76	0.77	1.96
Logistic Regression	0.64	0.60	0.6	2.62
Random Forest	0.64	0.61	0.61	1.66
KNN	0.68	0.66	0.66	1.44
XGBoost	0.59	0.57	0.57	9.90

Individual Dataset For Lung (*ID-L*)

Model	Accuracy	Balanced Accuracy	F1 Score	Time Taken
Linear SVC	0.97	0.96	0.97	1.63
Logistic Regression	0.97	0.96	0.97	2.95
Random Forest	0.93	0.93	0.93	2.33
KNN	0.76	0.75	0.76	1.70
XGBoost	0.93	0.93	0.93	7.96

Merged Dataset (*MD*)

Model	Accuracy	Balanced Accuracy	F1 Score	Time Taken
Linear SVC	0.82	0.80	0.82	16.62
Logistic Regression	0.80	0.75	0.79	11.11
Random Forest	0.80	0.77	0.79	2.37
KNN	0.68	0.63	0.68	2.71
XGBoost	0.80	0.77	0.79	52.04

3) Parameter Tuning: After finding out the best collection of classifiers for our dataset, the next step is to tune the parameters to obtain better results from them. For this purpose two approaches are going to be taken for this paper:

- K-fold cross-validation
- Randomized search

For a limited amount of samples, cross-validation can come in handy. After randomly shuffling the dataset, it is split into K groups. Then, one group is held out for testing purpose and the rest used as training dataset. Then the results are summarized with the means of the model skill scores. For randomized search, the goal is to find the hyperparameters that give the best result and only pass the set with these few hyperparameters.

D. Machine Learning Classifiers

1) Linear SVC Classifier: For parametric classification, each class needs to be distributed or typical feature space values need to be characterized properly. Support vector classifiers try to find the best hyperplane to separate the different classes. It does this by increasing the distance between sample points and hyperplane [10]–[15]. For our purpose, instead of using the linear kernel of SVC, the LinearSVC based on liblinear library is going to be used.

2) Logistic Regression Classifier: Logistic regression is a supervised learning method that is used to predict the probability of target variable. Using a sigmoid function, logistic regression can be explained. The function in Eqn. 1 takes an input x and outputs a probability value between 0 and 1.

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{1}$$

$$t = \beta_0 + \beta_1 x \tag{2}$$

Here, in equation 2, β_0 is the intercept and β_1 is the slope of the linear function t.

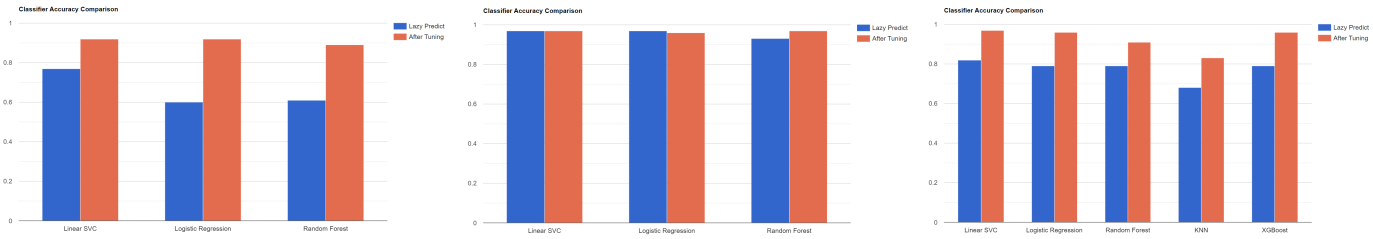
3) Random Forest Classifier: Random forests is an ensemble learning method for classification. It uses several decision trees on different subsets of the input dataset in order to increase the predicted accuracy and control over-fitting [16]–[21]. Here, each node of in the decision tree works on random subset of features to calculate the output. The random forest then aggregates the individual outputs that results in the final output. This intricate process is quite different from simple decision trees as they will be individually less accurate due to reduced amount of data but in an ensemble a better result will be produced due to reduced correlation between trees. Since, there are a lot of trees present in random forest, hence no individual trees need to be pruned.

4) KNN Classifier: K-nearest neighbor algorithm is based on supervised learning method. Here, for each new data, the similarity is between this and all available data and then then it is put into the category most similar to available categories. Rather than learning from the dataset immediately, it stores the new data and only at the time of classification does it perform any action. Initially, after choosing the value for K count of neighbors, the Euclidean distance was measured between K nearest neighbors and the closest K neighbors were chosen based on distance value. Then for each category, the number of data points among these K neighbors were counted. In the end, the new data point is assigned a category where the neighbor count is maximum.

5) XGBoost Classifier: Extreme gradient boosting is a specialized version of gradient boosting algorithm. It follows a level-wise strategy, scanning across gradient values and using the partial sums to evaluate the quality of splits at every possible split in the training set. Here, decision trees are created parallely and weights after being assigned to independent variables, are fed to the decision trees to predict results. Wrong predictions for a tree will see an increased weight and vice versa for the second decision tree. In the end, classifiers then ensemble to give a strong and more precise model.

TABLE III: Accuracy scores obtained from different classifiers

Dataset	Model	Class	Precision	Recall	f1-score	accuracy
ID-B	Linear SVC	0.0	0.94	0.92	0.93	0.92
		1.0	0.89	0.92	0.91	
	Logistic Regression	0.0	0.94	0.92	0.93	0.92
		1.0	0.89	0.92	0.91	
	Random Forest	0.0	0.88	0.94	0.91	0.89
		1.0	0.91	0.84	0.87	
ID-L	Linear SVC	0.0	0.97	0.98	0.98	0.97
		1.0	0.96	0.98	0.97	
	Logistic Regression	0.0	0.98	0.97	0.98	0.96
		1.0	0.97	0.96	0.96	
	Random Forest	0.0	0.98	0.96	0.97	0.97
		1.0	0.97	0.98	0.97	
MD	Linear SVC	0.0	0.97	0.98	0.98	0.97
		1.0	0.96	0.96	0.96	
		2.0	0.96	0.97	0.96	
	Logistic Regression	0.0	0.98	0.98	0.98	0.96
		1.0	0.96	0.97	0.96	
		2.0	0.96	0.94	0.95	
	Random Forest	0.0	0.96	0.98	0.97	0.91
		1.0	0.96	0.85	0.91	
		2.0	0.79	0.96	0.87	
	KNN	0.0	0.92	0.88	0.90	0.83
		1.0	0.82	0.84	0.83	
		2.0	0.77	0.77	0.77	
	XGBoost	0.0	0.98	0.98	0.98	0.96
		1.0	0.96	0.96	0.96	
		2.0	0.92	0.94	0.93	



(a) Bladder Dataset (b) Lung Dataset (c) Merged Dataset

Fig. 2: Comparing classifiers based on accuracy

IV. Results and Discussion

A. Performance Metrics

Obviously, the model needs to meet some required criteria for it to be considered successful. For our purpose, confusion matrix is going to be considered. In a confusion matrix, the predicted classes are represented in the columns whereas, the rows show the real class [22]. It can be easily determined how accurate this model is using the raw data from the confusion matrix. The 2x2 matrix reports the number of true positives (TP), true negatives (TN) , false positives (FP) & false negatives (FN). From these values, accuracy, recall, precision and f1-score for the classifier can be easily calculated by following the equations 3, 4, 5 & 6.

$$Accuracy\ (AC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{4}$$

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{5}$$

$$F1\ Score\ (F) = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

B. Comparative Analysis

From Table III, it is clearly evident that, “Linear SVC” provides the most amount of accuracy for all three datasets. For the merged dataset (*MD*), XGBoost achieves the same level of accuracy as “Logistic Regression” and surpasses “Random Forest” classifier.

As it can be seen from Fig. 2, almost all of the classifiers see an increase in accuracy after parameter tuning. Moreover, it is quite evident from the graph in Fig. 2c that, Linear SVC provides the most accurate classifier for our purpose. Not only that, two extra classifiers are considered for the merged dataset. Among these, KNN proves to be worst classifier based on accuracy but XGBoost proves to be almost as accurate as Logistic Regression even after not making the initial selection.

V. Future Works and Conclusion

Lung cancer and bladder cancer can be causally linked, so distinguishing between lung and bladder cancer tissues is critical for accurate diagnosis. The goal of this study was to determine the best method for classifying these tissues based on gene expression analysis data. Despite the fact that the number of classifiers considered was reduced, the results obtained after parameter tuning show that the classifiers are improving in accuracy. From our result, it can be seen that, "Linear SVC" might be the best possible options out there showing an accuracy of 97% . On the other hand, KNN gives the least amount of accuracy of only 83%.

For the merged dataset, only 5 classifiers were considered in this paper. However, using an accuracy-based rating system, only three classifiers were initially chosen, and this initial selection did not include the XGBoost classifier. It is quite impressive to see it outperform the Random Forest classifier, which was one of the initial choices. As a result, different parameter tuning applied to different classifier models using a different dataset may result in a different order, which may include a classifier that was not initially considered. Regardless, the findings in this paper are quite satisfactory in terms of achieving high accuracy in distinguishing between lung and bladder cancer tissues.

References

[1] H. Kuper, P. Boffetta, and H.-O. Adami, "Tobacco use and cancer causation: association by tumour type," *Journal of internal medicine*, vol. 252, no. 3, pp. 206–224, 2002.

[2] S. A. Strope and J. E. Montie, "The causal role of cigarette smoking in bladder cancer initiation and progression, and the role of urologists in smoking cessation," *The Journal of urology*, vol. 180, no. 1, pp. 31–37, 2008.

[3] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, no. 4, pp. 243–268, 2003.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[5] J. Taylor, A. B. Weiner, B. Wang, A. V. Balar, G. D. Steinberg, and R. S. Matulewicz, "Lung metastases versus second primary lung cancers in patients with primary urothelial carcinoma of the bladder: A national population-based assessment," *Bladder Cancer*, vol. 7, pp. 347–354, 2021, 3. [Online]. Available: https://doi.org/10.3233/BLC-210008

[6] E. Blaveri, J. P. Simko, J. E. Korkola, J. L. Brewer, F. Baehner, K. Mehta, S. Devries, T. Koppie, S. Pejavar, P. Carroll, and F. M. Waldman, "Bladder cancer outcome and subtype classification by gene expression," *Clin Cancer Res*, vol. 11, no. 11, pp. 4044–4055, Jun. 2005.

[7] D. Abdullah, A. Mohsin Abdulazeez, and A. Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," *Qubahan Academic Journal*, vol. 1, pp. 141–149, 05 2021.

[8] J. Choudhury, F. B. Ashraf *et al.*, "An analysis of different distance-linkage methods for clustering gene expression data and observing pleiotropy: Empirical study," *JMIR Bioinformatics and Biotechnology*, vol. 3, no. 1, p. e30890, 2022.

[9] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research," *Journal of Computational Biology*, vol. 26, no. 4, pp. 376–386, 2019, pMID: 30789283. [Online]. Available: https://doi.org/10.1089/cmb.2018.0238

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] J. C.-W. Chan, C. Huang, and R. Defries, "Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 693–695, 2001.

[12] C. Huang, L. Davis, and J. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of remote sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[13] M. Pal, "Support vector machine-based feature selection for land cover classification: a case study with dais hyperspectral data," *International Journal of Remote Sensing*, vol. 27, no. 14, pp. 2877–2894, 2006.

[14] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[15] M. Pal and G. M. Foody, "Evaluation of svm, rvm and smlr for accurate image classification with limited ground data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 5, pp. 1344–1355, 2012.

[16] G. Camps-Valls, L. Bruzzone, J. L. Rojo-Álvarez, and F. Melgani, "Robust support vector regression for biophysical variable estimation from remotely sensed images," *IEEE Geoscience and remote sensing letters*, vol. 3, no. 3, pp. 339–343, 2006.

[17] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[18] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *International journal of remote sensing*, vol. 26, no. 5, pp. 1007–1011, 2005.

[19] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[20] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24–31, 2016.

[21] Y. He, E. Lee, and T. A. Warner, "A time series of annual land use and land cover maps of china from 1982 to 2013 generated using avhrr gimms ndvi3g data," *Remote Sensing of Environment*, vol. 199, pp. 201–217, 2017.

[22] D. K. B. S. A. Janabi, "Data reduction techniques : A comparative study for attribute selection methods," 2018.