Predicting Employee Attrition using Machine Learning

Sarah S. Alduayj
School of Computer Science
University of Birmingham
Birmingham, United Kingdom
Email: sxa1115@student.bham.ac.uk

Kashif Rajpoot
School of Computer Science
University of Birmingham
Birmingham, United Kingdom
Email: k.m.rajpoot@cs.bham.ac.uk

Abstract— **The growing interest in machine learning among business leaders and decision makers demands that researchers explore its use within business organisations. One of the major issues facing business leaders within companies is the loss of talented employees. This research studies employee attrition using machine learning models. Using a synthetic data created by IBM Watson, three main experiments were conducted to predict employee attrition. The first experiment involved training the original class-imbalanced dataset with the following machine learning models: support victor machine (SVM) with several kernel functions, random forest and K-nearest neighbour (KNN). The second experiment focused on using adaptive synthetic (ADASYN) approach to overcome class imbalance, then retraining on the new dataset using the abovementioned machine learning models. The third experiment involved using manual undersampling of the data to balance between classes. As a result, training an ADASYN-balanced dataset with KNN (K = 3) achieved the highest performance, with 0.93 F1-score. Finally, by using feature selection and random forest, F1-score of 0.909 was achieved using 12 features out of a total of 29 features.**

Keywords— Machine learning; Employee attrition; Support vector machine; random forest; K nearest neighbours; Feature ranking; Feature selection.

I. Introduction

Employee attrition can be defined as the loss of employees due to any of the following reasons: personal reasons, low job satisfaction, low salary, and a bad business environment. Employee attrition can be categorised into two categories: voluntary and involuntary attrition. Involuntary attrition occurs when employees are terminated by their employer for different reasons, such as low employee performance or business requirements. In voluntary attrition, on the other hand, high-performing employees decide to leave the company of their own volition despite the company's attempt to retain them. Voluntary attrition can result from early retirement or job offers from other firms, for example. Although companies that realise the importance of their employees usually invest in their workforce by providing substantial training and a great working environment, they too suffer from voluntary attrition and the loss of talented employees. Another issue, hiring replacements, imposes high costs on the company, including the cost of interviewing, hiring and training.

Predicting employees attrition at a company will help management act faster by enhancing their internal policies and strategies. Where talented employees with a risk of leaving can be offered several propositions, such as a salary increase or proper training, to reduce their likelihood of leaving. Using machine learning models can help companies predict employees attrition. Using the historical data kept in human resources (HR) departments, analysts can build and train a machine learning model that can predict the employees who are leaving the company. Such models are trained to examine the correlation between the features of both active and terminated employees.

II. Related Work

Voluntary employee attrition is one of the major concerns for any company due to the severity of its impact. Talented employees are a major factor in business success and replacing such talent can be difficult and time-consuming [1]. Researchers have studied voluntary employee attrition and the factors responsible for it. The literature review shows that several factors can strongly contribute to employee attrition. For instance, [2] and [3] found that offering compensation is an important factor determining employee attrition as well as performance. The better the compensation, the lower the attrition rate. Whereas [1] found that money is not the only factor, as other combinations of factors, such as work load, performance pay and a weak career plan, have increased the attrition rate in the retail industry.

Several studies have explored the use of machine learning to predict employee behaviour. In [4], the authors used decision trees (ID3 C4.5) and Naïve Bayes classifier to predict employee performance. They found that job title was the strongest feature, whereas age did not show any clear effect. In [5] , the authors explored several data mining algorithms to predict employee churn (or attrition) using a dataset comprising 1575 records and 25 features. They used the following machine learning algorithms: naïve Bayes, support vector machines, logistic regression, decision trees and random forests. The research results recommend using a support vector machine (SVM), which has 84.12% accuracy. In [6], different decision tree algorithms were explored, including C4.5, C5, REPTree, and classification and regression trees (CART). The researchers tested and trained the decision trees using a dataset with total of 309 employee records out of 4326 records and a total of six features. As a result, the C5 decision tree gave the highest accuracy, at 74% compared with other decision trees. Also, their results showed that employee salary and length of service were important features in the tested organisation's dataset. Authors in [7] used neural networks to predict the turnover rate for small-west manufacturing company. Consequently, they developed the neural network simultaneous optimization algorithm (NNSOA) alongside 10-fold cross validation, which predicted the turnover rate with 94% accuracy. Moreover, they were able to identify the most important, relevant 'Tenure of employee on January 1' by using a modified genetic algorithm. In [8], a total of 6,909,746 employees' profiles available on the web were used to predict employee attrition. The employees' profiles included work experiences and education information along

978-1-5386-6673-9/18/$31.00 ©2018 IEEE

with company information. The researcher was able to train and evaluate an SVM model. The model prediction had 55% average accuracy, which is obviously not very high. The researcher recommended adding more personal features to the dataset, such as employee age, gender and work environment, which could improve the trained model. [9] predicted employee turnover for a global retailer located within the US. The dataset had 33 features and 73,115 observations. The researchers evaluated seven machine learning algorithms and found that XGBoost was the most accurate model, with a 0.88 area under the curve (AUC). In addition, it outperformed the other models with respect to memory utilization. In [10], the author developed a predictive model for employee attrition for Swedbank. In this study, a random forest model outperformed SVM and multi-layer perceptron (MLP) models with 98.6% accuracy.

Previous studies presented different accuracy measures where they used different machine learning models and various datasets. As a result, it is difficult to conclude which model is the best to use. In addition, previous studies didn't tackle the class imbalance problem which exists in real world attrition data. Therefore, we explored several methods to solve class imbalance which significantly enhanced the training process.

The remainder of this paper is organized as follows. Section 3 presents the proposed methods used in this research. Section 4 presents the experimental setup and results, and finally Section 5 concludes the research.

III. Proposed Methods

In this research, we have explored three main experiments to predict employee attrition. First, we have attempted to predict employee attrition using the original imbalanced dataset (data details presented in section IV). In the second experiment, we have introduced the adaptive synthetic sampling approach to solve the class imbalance problem. This approach involved oversampling the minority class which was in this case the "yes" class. The third experiment involved random under sampling of the data where we have randomly selected an equal subset of each class. Moreover, each experiment involved training and validating a set of machine learning classifiers to predict unseen dataset of employee attrition. All classifiers were validated using 5-fold cross validation. In addition, we have introduced feature selection method to minimize the trained models complexity and enhance their performance. In each case, each classifier was trained and evaluated iteratively by increasing the number of features for each iteration. The proposed methodologies are presented below with further details.

A. Classification

In this paper we have used several existing machine learning classification models to classify unseen data, and below we will introduce the classifiers used in this research.

Support vector machines (SVMs) is a non-probabilistic supervised machine learning model used for classification and regression. SVMs will train algorithms with assigned classes by separating each class through a decision boundary, also known as a hyperplane [11], [12].

Some problems are considered nonlinear in so far as it is difficult to draw the decision boundary. However, this can be solved by using a kernel function (also known as a kernel trick). This function returns the dot product of the two vectors, where it then maps data points to a new, transformed, high-dimensional space. Moreover, there are several types of kernel function can be used such as linear, Gaussian, and polynomial kernel [13], [14].

Random forest (RF) is one of the most powerful supervised machine learning algorithms for generating classifications and regressions. RF uses multiple decision trees to train data [15]. Each tree votes for a classification label for a certain dataset, then the RF model chooses which class had the most votes from the decision trees [16].

K-nearest neighbours (KNN) is one of the simplest machine learning algorithms and is used for both classification and regression. KNN works by specifying the value of K, which indicates the number of closest training points for a single data point. Each new data point will be classified based on the majority of votes collected from its neighbours [13] .

B. Adaptive synthetic (ADASYN) sampling approach

The ADASYN algorithm solves the class imbalance problem by creating new synthetic instances based on the density distribution of the minority class [20]. ADASYN will accomplish this by using adaptive learning to change the weights for the minority class instances. As a result, it will shift the decision boundary, which will make it easier to learn from difficult instances.

C. Feature Selection

Real-world datasets may include a large number of features. Some of these features are considered noise and might not have a positive influence on training machine learning algorithms. Using all available features will increase model complexity, hence affecting model performance and training time [21].

There are different methods that can be used to evaluate and rank all features. In this research, we use the t-test method which will calculate the mean and standard deviations for binary class labels used in the training data points. The t-test formula can be represented as following [22]:

$$t\,(x)=(\bar{y}_1(x)-\bar{y}_2(x))/\sqrt{(s_1^2(x)/n_1+s_2^2(x)/n_2)} \qquad (1)$$

where $\bar{y}_1(x)$ and $\bar{y}_2(x)$ are means for each class, while $s_1^2(x)$ and $s_2^2(x)$ are the standard deviations for class labels divided over number of samples n_1 and n_2.

IV. Dataset and Tools

The In this research, we used a publicly accessible dataset, which can be obtained from IBM Watson Analytics[1]. The dataset comprises synthetic data created by IBM data scientists. The dataset contains the HR-related data of 1470 employees with 32 features. Moreover, total of 1233 active employees were from "No" attrition category whereas the

[1] https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

remaining 237 former employees were from “Yes” attrition category In this research, two features were removed: ‘Employee count’, because it is a sequence of numbers (1,2, 3..) ; and ‘Standard hours’, since all employees have the same standard hours. Also, all non-numerical values were assigned numerical values for processing such as: Sales=1, Research & Development=2, Human Resources=3. Furthermore, MATLAB R2017b was used in this research to train and evaluate the machine learning models.

V. Experiments

In this section we illustrate the results of three main experiments performed on the dataset. Each experiment trained several machine learning models. All models were evaluated based on their accuracy, precision, recall, and F1 score. Further details are discussed in the below subsections.

A. *Performance Evaluation*

All trained models were evaluated by measuring their accuracy, precision, recall, and F1 score, which are described below: [17] [18] [19].

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FN+FP} \quad (2)$$

$$\text{Precision}= \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1 Score} = 2 * \frac{\text{Precesion} * \text{Recall}}{\text{Precesion}+\text{Recall}} \quad (5)$$

where TP are true positives values, TN are true negative, FP are false positives, and FN are the false negatives values.

B. *Imbalanced Data Experiments*

This section predicts employee attrition using the original class-imbalanced dataset. In this research, SVM, random forest, and KNN classification models were evaluated. First, each classifier was evaluated based on using all features in the dataset. Next, classifiers were evaluated by ranking and selecting the important subset features only.

Table I compares the performance of several classification models. Training with linear SVM yielded 86.9% accuracy but a very low F1 score. This indicates that it is misclassifying most of the minority class ‘Yes’. For further investigation, SVM was trained using different kernel types, such as quadratic, cubic and Gaussian. But, the F1 results were still low. The highest F1 score was 0.503, generated by quadratic SVM. The same applies to random forest and KNN. Although KNN was trained using different K values (1,3,5 and 10), the results were lower than those for SVM.

By using feature ranking it was possible to rank all features in the imbalanced dataset, where it first computed two-sample t-test and then returned an ordered index of the most important features which played a role in the training process. The top three of which were overtime, monthly income and job level. To further investigate, a linear SVM algorithm was trained and tested using only the top two features (overtime, monthly income), resulting in 83.9% accuracy. However, it did not classify any data point as ‘Yes’ for attrition; hence, F1 scored zero. Whereas when using SVM with several kernel types, the F1 scores slightly increased. Similar results were found using all three top features. Furthermore, training random forest with two features delivered a low F1 score. However, it increased in accuracy when using the top three features (overtime, monthly income, job level), yet still had very low F1 scores. In addition, KNN was trained with the top two and three features. In both experiments, KNN showed zero F1 results when K =1 and 5. Also, KNN results were significantly low when K = 3. In this research, feature selection continued up to 12 features, but the results were insignificant. As a result, using feature selection with imbalanced data did not show any significant improvement on model performance.

TABLE I Classifier performance with imbalanced data

Model Type	Accuracy	Precision	Recall	F1 Score[a]
Linear SVM	0.869	0.814	0.240	0.371
Quadratic SVM	0.871	0.662	0.405	**0.503**
Cubic SVM	0.841	0.508	0.418	0.458
Gaussian SVM	0.865	0.788	0.219	0.343
Random Forest	0.856	0.75	0.164	0.269
KNN (K=1)	0.827	0.275	0.046	0.079
KNN (K=3)	0.8374	0.25	0.004	0.008

[a] Bold values indicate highest F1 score

C. *Balancing Data Using Oversampling*

This section predicts employees attrition using a synthetically balanced dataset. This step was achieved by first scaling the dataset and then using ADASYN method. As a result, new synthetic data points were generated to oversample the minority class ‘Yes’. Therefore, the total number of observations in class ‘Yes’ increased up to 1152 observations, whereas class ‘No’ observations, comprising 1233 observations, did not change.

Table II compares the performance of several classification models when trained with all features, whereby the overall performance for all predictive models was significantly enhanced when trained with balanced classes. Training with linear SVM enhanced the F1 score up to 0.779. However, F1 scores were even higher when SVM was trained using quadratic, cubic and Gaussian kernels: the quadratic SVM had 0.881 F1 scores, the cubic SVM achieved 0.927 F1 scores, and the Gaussian SVM yielded 0.912 F1 scores. This indicates that the new balanced dataset is nonlinearly separable, and that using kernels to move data to higher dimensions helps to define the optimal boundary.

In addition, the balanced dataset was trained and evaluated using random forest. Unlike in the imbalanced dataset, random forest achieved 0.921 F1 scores. Furthermore, KNN was trained with several K values (1,3,5 and 10). KNN scored very high results when K = 1, which may indicate overfitting. Meanwhile, KNN achieved 0.931 and .909 F1 scores with K =3 and K = 5, respectively. Finally, KNN scored lower results with K = 10, where it achieved 0.88 F1 scores.

TABLE II. CLASSIFIER PERFORMANCE WITH SYNTHETIC BALANCED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[b]
Linear SVM	0.782	0.763	0.795	0.779
Quadratic SVM	0.879	0.839	0.927	0.881
Cubic SVM	0.926	0.879	0.981	**0.927**
Gaussian SVM	0.912	0.885	0.941	**0.912**
Random Forest	0.926	0.950	0.893	**0.921**
KNN (K=1)	0.967	0.939	0.997	**0.967**
KNN (K=3)	0.929	0.877	0.992	**0.931**
KNN (K=5)	0.904	0.843	0.987	**0.909**
KNN (K=10)	0.872	0.804	0.970	0.880

[b] Bold values indicate highest F1 score

After generating the synthetic data points, the feature ranking function was used to rank the top features that were contributing to the training process. As a result, it was found that the top three features were overtime, total working years and job level. Random forest scored the highest results compared with the other models, achieving a 0.829 F1 score, as shown in Table III. The remaining predictive models scored very low performance results when trained with the two features. Similar results were found when trained with only three features, whereas random forest reached 0.806 with only three features. The experiments continued in order to include the 12 top features. Table IV lists the top 12 features used in the training. Random forest was able to reach 0.909 F1 scores using only 12 subset features. Moreover, KNN with K = 3, 5 and 10 was able to score up to 0.882, .861 and 0.839, respectively. In addition, cubic and Gaussian SVMs reached more than 0.83 F1 scores.

D. Balancing Data Using Undersampling

In this section, we predict employee attrition using manual undersampling of the dataset to overcome class imbalance. This was done by randomly selecting an equal number of observations for each class, where each class had 237 observations. The new dataset had 474 total observations.

Table V compares the performance of several classification models when trained with all features. The maximum F1 score was reached via SVM, where the quadratic SVM score was 0.74, and both linear and Gaussian SVM scored 0.73. Moreover, cubic SVM and random forest reached 0.69 F1 scores. Finally, KNN had low results up to 0.59 when K = 10. These results indicate that using manual undersampling may lead to the loss of important information that may play a role in predicting attrition.

TABLE III. CLASSIFIER PERFORMANCE WITH FEATURE SELECTION FOR SYNTHETIC BALANCED DATA

Model Type	No. Features	Accuracy	Precision	Recall	F1 Score[c]
Linear SVM	2	0.648	0.676	0.523	0.589
Cubic SVM	2	0.593	0.550	0.871	0.674
Gaussian SVM	2	0.722	0.755	0.628	0.686
Random Forest	2	0.852	0.935	0.745	**0.829**
KNN (K=1)	2	0.659	1	0.294	0.454
KNN (K=3)	2	0.537	1	0.045	0.087
KNN (K=5)	2	0.523	1	0.016	0.032
Linear SVM	3	0.649	0.676	0.523	0.590
Cubic SVM	3	0.562	0.530	0.823	0.645
Gaussian SVM	3	0.722	0.753	0.630	0.686
Random Forest	3	0.826	0.869	0.752	**0.806**
KNN (K=1)	3	66.4	1	0.303	0.466
KNN (K=3)	3	0.572	0.995	0.175	0.298
KNN (K=5)	3	0.553	1	0.137	0.242
Cubic Linear	12	0.74	0.736	0.721	0.729
Cubic SVM	12	0.851	0.875	0.825	**0.850**
Quadratic SVM	12	0.801	0.796	0.791	0.794
Gaussian SVM	12	0.834	0.812	0.853	**0.832**
Random Forest	12	0.914	0.925	0.893	**0.909**
KNN (K=1)	12	0.641	1	0.256	0.407
KNN (K=3)	12	0.869	0.802	0.979	**0.882**
KNN (K=5)	12	0.844	0.771	0.976	**0.861**
KNN (K=10)	12	0.818	0.749	0.955	**0.839**

[c] Bold values indicate highest F1 score

TABLE IV. LIST OF TOP 12 FEATURES IN SYNTHETIC BALANCED DATA

1. Overtime	7. Stock Option Level
2. Total Working Years	8. Business Travel
3. Job Level	9. Job Role
4. Monthly Income	10. Job Involvement
5. Marital Status	11. Job Satisfaction
6. Years with Current Manager	12. Environment Satisfaction

TABLE V. CLASSIFIER PERFORMANCE FOR UNDERSAMPLED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[d]
Linear SVM	0.745	0.754	0.725	**0.739**
Quadratic SVM	0.747	0.760	0.722	**0.740**
Cubic SVM	0.707	0.733	0.650	0.689
Gaussian SVM	0.751	0.779	0.700	**0.738**
Random Forest	0.717	0.756	0.641	0.694
KNN (K=1)	0.589	0.595	0.552	0.573
KNN (K=3)	0.573	0.572	0.586	0.579
KNN (K=5)	0.565	0.562	0.586	0.574
KNN (K=10)	0.588	0.584	0.611	0.597

[d] Bold values indicate highest F1 score

Although undersampling results were low, feature ranking and selection were applied in this section. The feature ranking function was used to rank the top features that were contributing to the training process. As a result, it was found that the top three features were overtime, years with current manager, and total working years. Table VI shows the performance for all predictive models when using feature selection. Gaussian SVM, random forest and KNN all got very close results, between 0.66 and 0.68. In fact, KNN classified most observations as ‘Yes’. Furthermore, very close results were found when the three features were used during training.

TABLE VI. CLASSIFIES PERFORMANCE WITH FEATURE SELECTION FOR UNDERSAMPLED DATA

Model Type	No. Fea-tures	Accuracy	Precis-ion	Recall	F1 Score[e]
Linear SVM	2	0.652	0.698	0.536	0.606
Cubic SVM	2	0.631	0.661	0.536	0.592
Gaussian SVM	2	0.681	0.676	0.696	**0.686**
Random Forest	2	0.679	0.682	0.671	**0.677**
KNN (K=1)	2	0.523	0.511	0.995	**0.676**
KNN (K=3)	2	0.506	0.503	0.995	**0.668**
KNN (K=5)	2	0.5	0.5	1	**0.666**
Linear SVM	3	0.652	0.698	0.536	0.606
Cubic SVM	3	0.515	0.524	0.316	0.395
Gaussian SVM	3	0.67%	0.687	0.620	**0.652**
Random Forest	3	0.618	0.612	0.646	0.628
KNN (K=1)	3	0.5253	0.513	0.953	**0.667**
KNN (K=3)	3	0.5464	0.525	0.945	**0.675**
KNN (K=5)	3	0.5	0.527	0.919	**0.670**

[e] Bold values indicate highest F1 score

VI. CONCLUSION

A high employee attrition rate is a major problem for companies. Losing high-performing employees is considered a major loss for companies, specifically those that invest in their employees. Finding replacements with a similar level of performance is considered difficult and can cost the company both money and time.

The main objective of this research was to use machine learning models to predict employee attrition based on their features. This will give company management signs supported by machine learning tools. As a result, this will help management to act faster to reduce the likelihood of talented employees leaving their company. In this research, three experimental approaches were used on the dataset to develop predictive models. First, the original imbalanced data were trained via several predictive models, whereby quadratic SVM scored the highest results, with 0.50 F1 scores. Second, using the ADASYN approach, it was possible to balance between the two classes. It was noteworthy how the performance of all the models increased significantly: cubic, Gaussian, random forest and KNN (K = 3) achieved high F1 scores, between 0.91 and 0.93. Furthermore, very close results were achieved when using the feature selection: random forest achieved 0.92 F1 scores with two features only, and reached 0.90 using the top 12 features. The final technique was manually undersampling the dataset to have equal classes. As a result, important information was not captured and led to lower performance. Nevertheless, SVMs were able to capture more than 0.70 using all the features, and more than 0.60 with just two features.

VII. References

[1] S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry," Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.

[2] D. G. Gardner, L. V. Dyne and J. L. Pierce, "The effects of pay level on organization-based self-esteem and performance: a field study," Journal of Occupational and Organizational Psychology, vol. 77, no. 3, pp. 307-322, 2004.

[3] E. Moncarz, J. Zhao and C. Kay, "An exploratory study of US lodging properties' organizational practices on employee turnover and retention," International Journal of Contemporary Hospitality Management, vol. 21, no. 4, pp. 437-458, 2009.

[4] Q. A. Al-Radaideh and E. A. Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," nternational Journal of Advanced Computer Science and Applications, vol. 3, no. 2, p. 144–151 , 2012.

[5] G. K. P. V. Vijaya Saradhi, "Employee churn prediction," Expert Systems with Applications,, vol. 38, no. 3, pp. 1999-2006, 2011.

[6] D. A. B. A. Alao, "Analyzing employee attrition using decision tree algorithms," Computing, Information Systems, Development Informatics and Allied Research Journal, no. 4, 2013.

[7] R. S. Sexton, S. McMurtrey, J. O. Michalopoulos and A. M. Smith, "Employee turnover: a neural network solution," Computers & Operations Research, vol. 32, no. 10, pp. 2635-2651, 2005.

[8] Z. Ö. KISAO˘GLU, Employee Turnover Prediction Using Machine Learning Based Methods (Thesis), MIDDLE EAST TECHNICAL UNIVERSITY, 2014.

[9] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms," International

Journal of Advanced Research in Artificial Intelligence, vol. 5, no. 9, 2016.

[10] M. Maisuradze, Predictive Analysis On The Example Of Employee Turnover (Master's thesis), Tallinn: Tallinn University of Technology, 2017.

[11] K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," International workshop on multiple classifier systems, 2005.

[12] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?," Acm Sigkdd Explorations Newsletter, vol. 2, no. 2, pp. 1-13, 2000.

[13] S. Rogers and M. Girolami, A first course in machine learning, CRC Press, 2016.

[14] N. Cristianini and B. Scholkopf, "Support vector machines and kernel methods: the new generation of learning machines," Ai Magazine, vol. 23, no. 3, p. 31, 2002.

[15] T. K. Ho, "Random decision forests," in proceedings of the third international conference on Document Analysis and Recognition, 1995.

[16] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[17] X. Zhu, Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities, Igi Global, 2007.

[18] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine, 2011.

[19] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[20] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in IEEE International Joint Conference on Neural Networks, 2008.

[21] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of machine learning research, vol. 3, pp. 1157-1182, 2003.

[22] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," Proceedings of the National Academy of Sciences, vol. 100, no. 25, pp. 14666-14671, 2003.