IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Analysis of Machine Learning Algorithm: SMOTE with SVM

Pallavi Ahirwar[1], Anjna Jayant Deen[2], Manish Kumar Ahirwar[3]
[1]Student, [23]Associate Professor
[123]Computer Science and Engineering
[123]University Institute of Technology, RGPV, Bhopal, Madhya Pradesh

Abstract: **Imbalanced datasets are common in many real-world applications, such as fraud detection, disease diagnosis, and anomaly detection, where the occurrence of the target event is rare or infrequent. In general machine learning algorithms are prone to classify for majority class. In dealing with imbalanced class datasets various problems occur, however, the majority class samples can be leading to a misleading impression of accuracy and desired target. In handling such a situation that can use by many researchers is SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples for the minority class and resolves the over-fitting problem caused by random over-sampling of the majority class. This study thoroughly investigated SMOTE to discover the noteworthy outcomes of classification algorithms for imbalanced datasets. This creates an oversampled dataset for minority class to try to reduce over fitting and unbefitting problems of class imbalance; and creates more balanced dataset for training in classification models. In this study SMOTE can be used with various classification algorithms, such as KNN, Decision trees, Random forests, and Support vector machine for analysis and betterments of classification algorithms.**

Indexed Terms: SMOTE, SVM, majority class, minority class, imbalanced data, lung cancer data.

INTRODUCTION

SMOTE (Synthetic Minority Over-sampling Technique) [18] and SVM (Support Vector Machine) [19] are two commonly used techniques in machine learning. SMOTE is a technique used to address the class imbalance problem in classification tasks. In a class-imbalanced dataset, one class may have significantly fewer examples than the other, which can lead to a model that is biased towards the majority class. SMOTE works by creating synthetic examples of the minority class by interpolating between existing examples. This can help to balance the classes and improve the accuracy of the model. SVM, on the other hand, is a type of supervised learning algorithm that is used for classification and regression analysis. SVM works by finding the optimal hyperplane that separates the data into two classes. Accurate segmentation of lung cancer in pathology slides is a critical step in improving patient care. [24] Lung cancer (LC) is one of the most serious cancers threatening human health. Histopathological examination is the gold standard for qualitative and clinical staging of lung tumors [25] The hyperplane is chosen such that it maximizes the margin between the two classes. SVM is particularly effective in high-dimensional spaces, and it can handle both linear and non-linear decision boundaries. In summary, SMOTE is a technique used to address class imbalance, while SVM is a classification algorithm that finds the optimal hyperplane to separate the data into two classes. Together, they can be used to improve the accuracy of classification models on imbalanced datasets. The SVM algorithm finds the optimal hyperplane that separates the data into two classes.

The equation of the hyperplane in a two-dimensional space is given by eq (1) to eq (2):
$w_0 + w_1 x_1 + w_2 x_2 = 0$……eq (1)
where x_1 and x_2 are the two features of the data, w_1 and w_2 are the coefficients of the hyperplane, and w_0 is the intercept.

In a high-dimensional space, the equation of the hyperplane is given by:
$w_0 + w_1 x_1 + w_2 x_2 + ... + w_n x_n = 0$……. eq (2)
where $x_1, x_2, ..., x_n$ are the features of the data, $w_1, w_2, ..., w_n$ are the coefficients of the hyperplane, and w_0 is the intercept.
The SVM algorithm finds the values of $w_0, w_1, w_2, ..., w_n$ that maximize the margin between the two classes. The margin is the distance between the hyper plane and the closest points of each class. The points that lie on the margin are called support vectors. Once the optimal hyper plane is found, the SVM algorithm can use it to classify new data points. If the value of the equation for a new data point is greater than 0, then the point is classified as belonging to one class, and if it is less than 0, then the point is classified as belonging to the other class. The learning method known as the support vector machine (SVM) was created in 1990. This approach is based on findings from Vapnik' s statistical learning theory [20]. [19] A key idea for the majority of learning problems, kernel functions, is closely related to SVM machines. SVM and the kernel framework are applied in many different disciplines. Lung cancer is known to be one of the most dangerous diseases which are the main reason for disease and death when

diagnosed in primitive stages. [28] Bioinformatics, pattern recognition, and multimodal information retrieval are all included. The machine learning method known as a support vector machine (SVM) examines data for regression and classification purposes. A supervised learning technique called SVM organizes data into one of two groups after looking at it. The sorted data are produced as a map by an SVM, [22] with the margins between the two being as far away as feasible. SVMs are employed in the sciences, picture classification, handwriting recognition, and text categorization. It is trained using a set of data that has already been divided into two categories, creating the model as it learns. Lung nodule classification plays an important role in diagnosis of lung cancer which is essential to patients' survival [26]. lung cancer is responsible for nearly one in five cancer deaths. The National Lung Screening Trial (NLST) demonstrated the efficacy of Low-Dose Computed Tomography (LDCT) [27] An SVM algorithm's job is to ascertain which category a new data point falls within. As a result, SVM may be considered a non-binary linear classifier. In addition to classifying items, an SVM method should maximize the space on a network between each object. SVM may be applied to classification or regression problems. However, categorization issues are where it is most frequently utilized. When using the SVM algorithm, each data point is represented as a point in n-dimensional space (where n is the number of features you have), with each feature's value being the value of a certain coordinate. Next, we do classification by identifying the hyper-plane that effectively distinguishes the two classes.

Proposed Methodology

The combination of SMOTE with SVM can be useful in dealing with imbalanced datasets where the minority class has significantly fewer samples than the majority class. SMOTE can help to generate synthetic samples of the minority class, which can be used to train an SVM classifier. T-link. This model helps to improve the performance of the SVM classifier on the minority class as shown in fig. (1) AND fig (2)

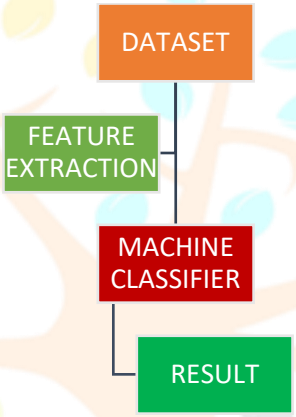


Fig 1 Flow chart of Proposed Methodology

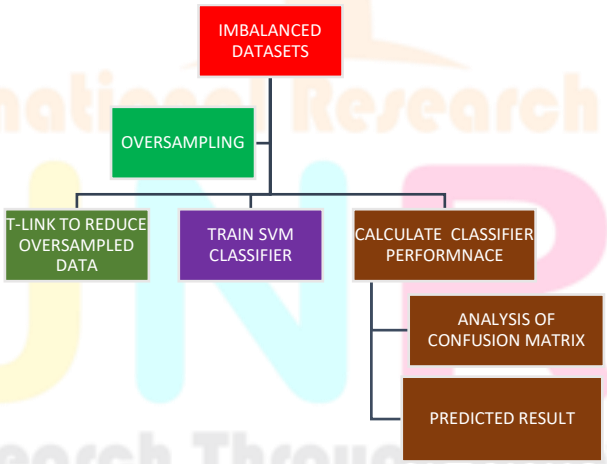


Fig 2 Proposed Methodology

LITERTURE SURVEY

Jae-Hyun Seo et.al. [1] created a model using support vector machine along with SMOTE. After randomly generating number of tuples of smote ratios, these tuples are used for creation of numerical model for optimizing SMOTE ratios of rare class. When it comes to false positives in intrusion detection (ID), an effective solution is presented utilizing SMOTE tuples. In order to solve the issue of class imbalance between normal class and attack classes, certain tuples of SMOTE ratios are formed, and these tuples are utilized to develop a model using an SVM (Support Vector regression). KDD cup dataset is used as an imbalanced dataset for intrusion detection. Accuracy of 98.1% is acquired by using SVM with SMOTE.

Jia-Bao Wang et.al [2] proposed SMOTE with SVM to solve the issue of class disparity in learning, it also introduces interpolation and co-linearity between the new points and the old ones. To address this issue, propose the adaptive-weighing SMOTE (AWSMOTE) method, which uses one type of weight in the variable space to address the drawbacks of co-linearity and another type of weight in the sample space to select SVMs from the minority class for the express purpose of using them as weight points

in interpolation. When compared to SMOTE, AWSMOTE operates more efficiently and produces more effective outcomes while also offering a number of other benefits. This method removes collinearity and uses SVM for weight points to remove interpolation. 1.2727 accuracy is attained by using AWSMOTE. six simulated datasets and twenty-two UCI and KEEL imbalanced datasets, metabonomic dataset.

Bo ZHOU et.al. [3] In the study, a combination approach of constructed SMOTE and quasi-linear SVM is suggested. An SVM with a quasi-linear kernel function is known as a quasi-linear SVM. By combining multiple local linear boundaries with interpolation, it creates an approximation of a nonlinear separation boundary. The constructed SMOTE prevents the possibility of class overlap by implementing oversampling while taking data distribution information into account. This experiment evaluates the performance of the constructed SMOTE with quasilinear SVM using the "yeast" datasets (Fun Cat 14). There are 436 characteristics and 962 individuals in the dataset. Accuracy attained by proposed method was 84.27 %.

Xin Wang et.al [4] proposed an oversampling method AGNES-SMOTE (Agglomerative Nesting-Synthetic Minority Oversampling Technique), which is based on enhanced SMOTE and hierarchical clustering and SVM used for imbalanced data. Its primary techniques involve clustering minority and majority samples hierarchically, splitting minority sub clusters based on obtained majority sub clusters, choosing a "seed sample" based on the sampling weight and probability distribution of the minority sub cluster, and limiting the generation of new samples in a specific area using the centroid method during the sampling process. The proposed method has higher accuracy than other compared methods and dataset was taken from UCI repository.

T. DEEPA et.al [5] give a new technique E-SMOTE Technique for balancing the dataset. Micro array dataset is used to assess and use SVM classification for choosing the features. E-SMOTE is based on genetic algorithm. Lymphoma and lung cancer micro array dataset are used for the investigation. The dataset has 96 characteristics and data in total which gives accuracy of 72% for lymphoma and 36% for lung cancer.

Andrew Christian Flores et.al [6] compare the support vector machine with SMOTE and Naïve Bayes with SMOTE for dataset of sentiment analysis. The Duterte Administration Tweets from Twitter are utilized as dataset A, while Impact of K-12 Program in the Philippines communications are used as dataset B, in Sentiment Analysis. Accuracy for SVM with SMOTE for dataset A and B was 82.88% and 84.68% respectively.

Qinghua Cao et.al [7] proposed a brand-new oversampling method called SMOBD (Synthetic Minority Over-sampling Based on samples Density*)* which performs better. Additionally, we mix this method with various error costs SVM. 9 UCI datasets were used to test the algorithm. The outcomes demonstrate that the SMOBD-CS (Synthetic Minority Over-sampling Based on samples Density and Cost-sensitive SVM) we suggested can enhance classifier performance for unbalanced datasets.

Qi Wang et .al [8] In order to address unbalanced data learning (IDL) issues, a novel ensemble approach dubbed Bagging of Extrapolation Borderline-SMOTE SVM (BEBS) has been developed. Experimental datasets are selected from the UCI Machine Learning Repository. For handling the IDL in binary situation, a brand-new ensemble method named BEBS was suggested. An adaptive sampling strategy, extrapolation borderline-SMOTE, and bootstrapping aggregation to the previously unbalanced dataset were used to frame the BEBS Proposed algorithm greatly outperformed other models in roughly 76.2% of comparison outcomes, which were calculated as the total number of paired comparisons.

Yuan Sui et.al. [9] a classifier SVM classifier is proposed with random under sampling RU and SMOTE to recognize lung nodules. For training and classification, eight features, comprising 2D and 3D features, are retrieved. The RU-SMOTE-SVM classifier has the greatest classification accuracy among the four types of classifiers, according to experimental results, and its average classification accuracy is higher than 92.94% for training datasets of various sizes. Low-dose CT lung scans from Sheng Jing Hospital, a part of Chinese Medical University, Beijing Xuanwu Hospital, and the U.S. National Cancer Institute (NCI) provided by the Lung Image Data Union are the experimental data used (Lung Image Database Consortium, LIDC) The proposed approach of combining random undersampling and SMOTE with SVM has shown promising results in several studies on lung nodule recognition. It has been found to improve the sensitivity and specificity of the classifier, leading to more accurate diagnoses. In summary, the combination of random undersampling and SMOTE with SVM is a promising approach for lung nodule recognition in medical imaging. This approach can help to address the imbalanced nature of the datasets and improve the accuracy of the classifier. Further research in this area could lead to even better results and more accurate diagnoses.

Muhammad Tahir et. al. [10] compare other current state-of-the-art approaches, the Golgi-predictor has demonstrated significant performance and obtained promising outcomes. The suggested system acquired an accuracy rating of 97.6% during the 10-fold cross-validation to anticipate any unusual activity of the Golgi apparatus. dataset comprising of 87 *cis*-Golgi and 217 *trans*-Golgi protein sequences originally constructed by **Yang et. al.** [11] The proposed approach in this paper involves using a hybrid feature space, which combines different types of features, such as amino acid composition, dipeptide composition, and physicochemical properties. This can help to capture the different aspects of the protein sequence and improve the discriminatory power of the classifier. In addition to the hybrid feature space, the authors also propose using an ensemble of support vector machines (SVMs) to classify the Golgi proteins. The ensemble approach involves combining multiple SVMs with different parameter settings, which can help to improve the overall accuracy and robustness of the classifier. In summary, the proposed approach in this paper is a promising method for discriminating Golgi proteins using machine learning. [23] The use of a hybrid feature space, SMOTE, and an ensemble of SVMs can help to address the imbalanced nature of the dataset and improve the accuracy and robustness of the classifier. Further research in this area could lead to even better results and more accurate characterizations of Golgi proteins.

Jiang Shen et. al. [12] A hybrid SMOTE and adaptive SVM approach is presented for analyzing survival data from lung cancer patients (LCPs) who have undergone surgery. The suggested strategy is broken down into two phases: SMOTE's performance may be enhanced by filtering out noisy samples with the use of the cross verified committee filter (CVCF). In the second phase, FPSO employs an adaptive support vector machine to fine-tune the SVM's underlying parameters. There is a 95.11% improvement in accuracy, a 95.10% improvement in the G mean, a 95.02% improvement in the F1, and an AUC of 95.01% for LCPs using this strategy.

Ratchakoon Pruengkarn et. al. [13] For solving the unbalanced classification issue, a hybrid sampling strategy is presented by merging the Complementary Fuzzy Support Vector Machine (CMTFSVM) and Synthetic Minority Oversampling Technique (SMOTE). The suggested method is evaluated against three other classifiers and employs an optimized membership function to improve classification performance. Four common benchmark datasets were used in the trials from KEEL AND UCI repository along with a real-world dataset of plant cells. The outcomes showed that using CMTFSVM followed by SMOTE produced superior results for the benchmark datasets compared to other FSVM classifiers. It also displayed the greatest results on real-world datasets, with G-mean and AUC values of 0.9589 and 0.9598, respectively.

Phakhawat Sarakit1 et.al. [14] employ a sampling-based technique called SMOTE to enhance the performance of emotion categorization by oversampling cases from minority classes to the number of instances from the majority class. The YouTube dataset was balanced using the SMOTE method and evaluated with the multinomial Naive Bayes (MNB), decision tree (DT), and support vector machines machine learning techniques (SVM). SVM gets the most accuracy, scoring 93.30% on the filtering job and 89.44% on the classification assignment. Two datasets of comments in Thai gathered from Thai YouTube videos were used the YouTube API 2.0 to randomly choose 85 video clips, one from each data set, from the total. 2,480 samples make up the emotion filtering dataset, which separates comments into three categories. While the Emotion Classification dataset has 5,345 samples with comments divided into six emotion categories.

Josey Mathew et. al [15] offers a kernel-based SMOTE (KSMOTE) technique that stably produces synthetic minority data points in the SVM classifier's feature space. By enhancing the initial Gram matrix based on neighborhood data in the feature space, additional data points are added. On 51 benchmark datasets, it has been statistically demonstrated that the suggested method performs better. In a semiconductor etching chamber, K-SMOTE is also used to anticipate the stage of deterioration, where it obtains a greater level of accuracy for the unbalanced defective stages. The KEEL data repository's 51 binary datasets are chosen to test the efficiency of the suggested method.

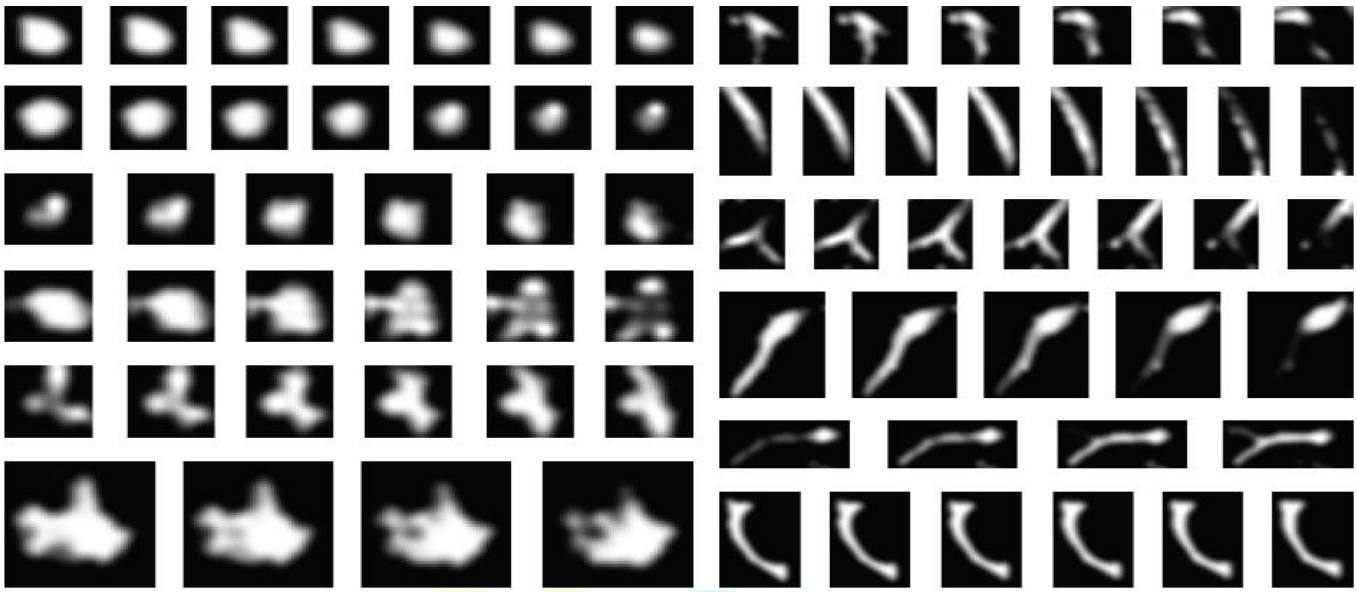
Liliya Demidova et. al. [16] In this research, a novel method for choosing the best parameter values for the SMOTE (Synthetic Minority Over-sampling Technique) algorithm in the classification of unbalanced datasets using SVM (Support Vector Machine) has been proposed. With this method, the amount of time needed to find the SMOTE algorithm's ideal parameter values is reduced. The results of the experiments demonstrate that the suggested strategy enables improving the SVM classifier's classification quality. The classification performance of the technique suggested in this research was demonstrated using actual medical datasets from the UCI repository of the machine learning database of heart and hepatitis. Accuracy of 95.67% and 92.90% is attained respectively.

RESULTS AND DISCUSSION

As discussed in article [9]. They are using all images were of the size 512 × 512 pixels. The pixel size varied from 0.488mm to 0.762mm, and the slice thickness ranged from 1.25mm to 3.0 mm. In this study we consider the same image pixel value in our implementation. We consider 100 thoracic CT scans for the experiments. As researcher they discussed in their article [9] the same parameter we are using here to extracted nodule and nonnodule regions from the lung images. Forming the nonnodule and nodules region data based on a, b and c layers value in x, y and z direction. The range of pixel value here we consider same as discussed in articles [9]. create 130 nodules and 808 nonnodules for the dataset as shown in Figure () in order to this k fold cross-validation randomly divided nodules and nonnodules training samples that is 50 and 430 respectively; k = 5.

Table 1 Data samples

Region of Interest DATASET	NUMBER OF NODULES	NUMBER OF NONNODULES
Original training samples	50	430
Balanced training samples by SMOTE method	380	465
SMOTE+SVM method	180	225
Testing data	65	415



(a) Nodules images **(b)Non nodules images**

Fig 3 Nodule and nonnodule sequent images. cc @[9]

Studies have shown that the combination of SMOTE with SVM can improve the accuracy and F1-score of the classifier, especially when the dataset is highly imbalanced. However, it is important to note that the effectiveness of this approach depends on various factors such as the choice of parameters, the quality of the dataset, and the specific problem at hand.

In summary, studying SMOTE with SVM can be useful for researchers and practitioners working with imbalanced datasets in various domains such as healthcare, finance, and security**.**

The experimental data used are low-dose CT lung images from Sheng Jing Hospital affiliated to Chinese Medical University, Beijing Xuanwu Hospital, and the U.S. National Cancer Institute (NCI) issued by the Lung Image Data Union (Lung Image Database Consortium, LIDC) [17] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

Table 2 Summary of recent work based on SMOTE with SVM

Author's Name	Algorithm Used	Dataset	Data Sample Size	Data features	Accuracy
Jae-Hyun et.al [1]	SMOTE + SVM	1999 KDD cup Dataset [86]	User to root (U2R), Remote to local (R2L), Denial of service (DoS), and Probe.	100 for each	98.1 %
Jia - Bao et.al. [2]	AWSMOTE + SVM	UCI, KEEL & Metabonomic [55]	Total samples – 100	No. of features 50 & 500	57.58
B. ZHOU et. al [3]	Quasi linear SVM + SMOTE	Yeast dataset (Fun cat 14)	962 individuals	463 features	84.27%
Xin Wang et.al. [4]	AGNES-SMOTE + SVM	UCI dataset [85]	9 samples are taken Ecoli, Libra, Yeast1, Optical digits,	7, 90, 10, etc….	85.83
T. DEEPA et. al. [5]	E-SMOTE + SVM	Micro array dataset of lymphoma and lung cancer.	96 data	Lymphoma – 43 Lung-52	Lymphoma – 72%
Andrew Christian Flores et.al [6]	SMOTE+ SVM & Naïve Bayes	Twitter & K12 program in Philippines [56], [57]	Total instances 2494 and 1051	2067 and 890	84.68%
Qinghua Cao et.al [7]	SMOBD + SVM	UCI datasets	9 samples-different characteristics	Glass, abalone, hypothyroid, hepatitis	AUC- 91.9
Qi Wang et .al [8]	Borderline-SMOTE + SVM	UCI [87]	6 samples	10,23,8,19	83.3%

Yuan Sui et.al. [9]	SVM +RU + SMOTE	Chinese Medical University	150 nodule and 908 nonnodule	75 nodules & 454 nonnodular	92.94%
Muhammad Tahir et. al. [10]	SMOTE + SVM	(UniProtKB) [25], [51], [52], [53]	Golgi proteins	87 *cis*-Golgi and 217 *trans*-Golgi protein	10-fold - 97.6%
Jiang Shen et. al. [12]	Hybrid SMOTE + adaptive SVM	Wroclaw Thoracic Surgery Centre [58]	470 samples	36 features	95.11%
RatchakoonPruengkarn et. al. [13]	SMOTE + SVM	KEEL, UCI, DNA cells of plants detection. [59], [60], [61]	80% training dataset & 20% testing dataset	5 features	95.98
Phakhawat Sarakit1 et.al. [14]	SMOTE + SVM Naïve Bayes, Decision Tree	YouTube dataset	Emotion filtering dataset 2,480 Emotion Classification dataset 5,345	5000 features for each sample	93.30%
Josey Mathew et. al [15]	KSMOTE + SVM	KEEL	51 binary datasets	90 to 2084	66 %
Liliya Demidova et. al. [16]	SMOTE + SVM	UCI [62]	**Dataset** / **ratio**: Heart 0.2; Hepatitis 0.74	2 features	Hepatitis 2-100%

Performance Matrix

In this study the performance evaluation mainly Recall eq. (3), Precision eq. (4), F1- score eq. (5) confusion matrix, and Accuracy eq. (6) are taken for consideration. A brief summary for these measures is discussed below:

Recall – Recall is a metrics that measures how many correct positive predictions were made out of all possible positive predictions as shown in eq. (3) below:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \ldots\ldots\ldots\ldots\ldots (3)$$

Precision – Precision is calculated by dividing the total number of true positive results by the sum of total true positives and false positives as shown in eq. (4) below:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \ldots\ldots\ldots\ldots (4)$$

F1- Score - An F-score is the harmonic mean of a system's precision and recall values. It can be calculated by the following formula as shown in eq. (5) below:

$$\text{F1- Score} = \frac{(2*Precion*Recall)}{(Precision+Recall)}) \qquad \ldots\ldots\ldots\ldots (5)$$

Accuracy -One parameter for assessing classification models is accuracy. The percentage of predictions that our model correctly predicted is known as accuracy. The following is the official definition of accuracy as shown in eq. (6) below:

$$\text{Accuracy} = \frac{TP+FN}{TP+FN+FP+FN} \qquad \ldots\ldots\ldots\ldots\ldots (6)$$

Confusion Matrix: A confusion matrix aids in visualizing the results of a classification task by providing a table arrangement of the various outcomes of the prediction and findings. It displays a table with all of a classifier's predicted and actual values as shown in table 3 below

:

Table 3 Confusion Matrix

	Actual negative	**Actual positive**
Predict negative	True negative (TN)	False negative (FN)
Predict positive	False positive (FP)	True positive (TP)

Table 4 Accuracy of Proposed classifier

Evaluation index classifier	TP	FN	FP	TN	Accuracy
SVM classifier	20	50	4	456	88.76
SVM classifier (RBF kernel)	55	35	27	440	90.02
SMOTE-SVM classifier (Proposed model)	55	28	19	436	93.25

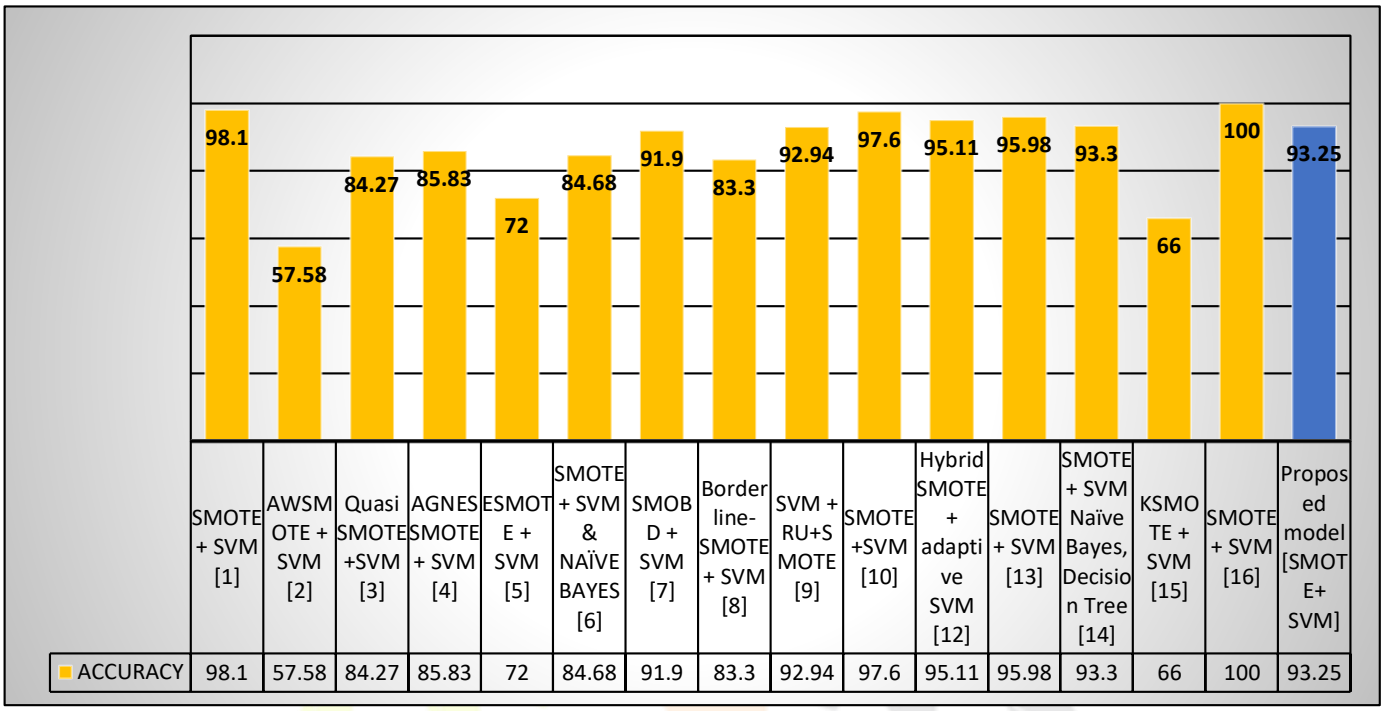


Fig 4 Accuracy bar Chart for SMOTE with SVM

As discussed datasets in table 1, number of nodules and nonnodules from original datasets are 50 and 430 respectively, and testing datasets is 65 and 415 for nodules and nonnodules respectively. Fig 4 displays a bar chart of an investigation of SMOTE + SVM accuracy using an unbalanced dataset. In the aforementioned research, all the authors accomplished great work in SMOTE under varied datasets. T-link uses for removing oversampled data to balance oversampling. Accuracy ranges from 57.58 to 100 for SMOTE with SVM classifier. In article [9] achieved accuracy is 92.94% and most of the research work has higher accuracy is achieved due to text data for our proposed work accuracy achieved is 93.25 %. In table 4, SVM kernel RBF (Radial Basis function) achieved accuracy is 90.02, In this study, Proposed classifier SMOTE + SVM accuracy successfully achieved 93.25%. For data samples. In summary, using SMOTE to oversample the minority class can be an effective way to address the class imbalance problem in SVM classification. However, it is important to fine-tune the parameters of both the SVM and SMOTE to achieve the best possible performance.

CONCLUSION

In conclusion, SMOTE is a useful technique for addressing the class imbalance problem in machine learning. It helps to improve the classification performance of a model by generating synthetic samples for the minority class, which increases its size and reduces bias towards the majority class. SMOTE is easy to implement and can be used with various machine learning algorithms and performance metrics. However, it's important to note that SMOTE may not always work well in all scenarios and its effectiveness should be evaluated on a case-by-case basis. Computer-Aided Lung Nodule Recognition is an important area of research in medical image analysis. One of the major challenges in this field is dealing with imbalanced datasets, where the number of positive (nodule) samples is much smaller than the negative (non-nodule) samples. In recent years, several studies have explored the use of machine learning algorithms, such as SVM, for the automatic detection of lung nodules from CT images. However, the imbalanced nature of the dataset can lead to poor performance of the classifier, especially on the minority class. To address this issue, this study proposed a combination of SVM and SMOTE techniques to balance the dataset and improve the performance of the SVM classifier the importance of feature selection in improving the performance of the classifier. In the past, researchers developed computer-aided diagnosis (CAD) systems, which were greatly used by the radiologist for identifying the abnormalities and applied few features extracting methods. [29] Non-small cell lung cancer (NSCLC) is the most prevalent form of lung cancer and a leading cause of cancer-related deaths worldwide. [30] The authors used a combination of geometric and texture features extracted from

the lung nodules, and found that the SVM classifier achieved the best performance when using a subset of the most relevant features. For future scope CNN model can help in achieving more higher accuracy and result.

REFRENCES

[1] Seo, J.-H. & Kim, Y.-H. Machine -learning approach to optimize smote ratio in class imbalance dataset for intrusion detection. Computational Intelligence and Neuroscience2018, 1–11 (2018).
[2] Wang, J.-B., Zou, C.-A. & Fu, G.-H. AWSMOTE: An SVM-based adaptive weighted smote for class-imbalance learning. Scientific Programming2021, 1–18 (2021).
[3] Zhou, B., Yang, C., Guo, H., & Hu, J. (2013). A quasi-linear SVM combined with assembled SMOTE for imbalanced data classification. The 2013 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn.2013.6707035
[4] Wang, X., Yang, Y., Chen, M., Wang, Q., Qin, Q., Jiang, H., & Wang, H. (2020b). AGNES-SMOTE: An Oversampling Algorithm Based on Hierarchical Clustering and Improved SMOTE. Scientific Programming, 2020, 1–9. https://doi.org/10.1155/2020/8837357
[5] Deepa, T., &Punithavalli, M. (2011). An E-SMOTE technique for feature selection in High-Dimensional Imbalanced Dataset. 2011 3rd International Conference on [Electronics Computer Technology. https://doi.org/10.1109/icectech.2011.5941710
[6] Flores, A. C., Icoy, R. I., Pena, C. F., &Gorro, K. D. (2018). An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set. 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST). https://doi.org/10.1109/iceast.2018.8434401
[7] Cao, Q., & Wang, S. (2011). Applying Over-sampling Technique Based on Data Density and Cost-sensitive SVM to Imbalanced Learning. 2011 International Conference on Information Management, Innovation Management and Industrial Engineering. https://doi.org/10.1109/iciii.2011.276
[8] Wang, Q., Luo, Z., Huang, J., Feng, Y., & Liu, Z. (2017). A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM. Computational Intelligence and Neuroscience, 2017, 1–11 https://doi.org/10.1155/2017/1827016
[9] Sui, Y., Wei, Y., & Zhao, D. (2015). Computer-aided lung nodule recognition by SVM classifier based on combination of random under sampling and smote. Computational and Mathematical Methods in Medicine, 2015, 1–13 https://doi.org/10.1155/2015/368674
[10] Tahir, M., Khan, F., Rahmani, M. K. I., & Hoang, V. T. (2020). Discrimination of Golgi Proteins Through Efficient Exploitation of Hybrid Feature Spaces Coupled with SMOTE and Ensemble of Support Vector Machine. IEEE Access, 8,206028–206038. https://doi.org/10.1109/access.2020.3037343
[11] R. Yang, C. Zhang, R. Gao, and L. Zhang, ``A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data," Int. J. Mol. Sci., vol. 17, no. 2, p. 218, Feb. 2016.
[12] Shen, J., Wu, J., Xu, M., Gan, D., An, B., & Liu, F. (2021). A Hybrid Method to Predict Postoperative Survival of Lung Cancer Using Improved SMOTE and Adaptive SVM. Computational and Mathematical Methods in Medicine, 2021, 1–15. https://doi.org/10.1155/2021/2213194
[13] Pruengkarn, R., Wong, K. W., & Fung, C. C. (2017). Imbalanced data classification using complementary fuzzy support vector machine techniques and SMOTE. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). https://doi.org/10.1109/smc.2017.8122737
[14] Sarakit, P., Theeramunkong, T., & Haruechaiyasak, C. (2015). Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA). https://doi.org/10.1109/icaicta.2015.7335373
[15] Mathew, J., Luo, M., Pang, C. K., & Chan, H. L. (2015). Kernel-based SMOTE for SVM classification of imbalanced datasets. IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society. https://doi.org/10.1109/iecon.2015.7392251
[16] Demidova, L., & Klyueva, I. (2017). SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem. 2017 6th Mediterranean Conference on Embedded Computing(MECO). https://doi.org/10.1109/meco.2017.7977136
[17] S. G. Armato III, G. McLennan, M. F. McNitt-Gray et al., "Lung image database consortium: developing a resource for the medical imaging research community," Radiology, vol. 232, no. 3, pp. 739–748, 2004.
[18] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegel Meyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research16 https://doi.org/10.1613/jair.953
[19] Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh (1992)
[20] V.N.Vapnik,*TheNature of Statistical LearningTheory*, Springer,New York, NY, USA, 2nd edition, 1998.
[21] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings*, vol. 3201 of *Lecture Notes in Computer Science*, pp. 39–50, Springer, Berlin,Germany, 2004.
[22] J. Tian, H. Gu, and W. Liu, "Imbalanced classification using support vector machine ensemble," *Neural Computing & Applications*, vol. 20, no. 2, pp. 203–209, 2011.
[23] V.G. Bram, M. H. Bart, and A.Max, "Computer-aided diagnosis in chest radiography," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1228–1241, 2001.
[24] Li, Z., Zhang, J., Tan, T., Teng, X., Sun, X., Zhao, H., Liu, L., Xiao, Y., Lee, B. J., Li, Y., Zhang, Q., Sun, S., Zheng, Y., Yan, J., Li, N., Hong, Y., Ko, J., Jung, H. S., Liu, Y., . . . Litjens, G. (2021). Deep Learning Methods for Lung Cancer Segmentation in Whole-Slide Histopathology Images—The ACDC@LungHP Challenge 2019. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/jbhi.2020.303974

[25] Minteer, S. D., Ma, X., Chen, C., Yuan, Y., Zhang, S., Yan, Z., Chen, F., Bai, Y., Zhou, P., & Ma, M. (2021). Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images. *IEEE Access*, *9*, 53687–53707. https://doi.org/10.1109/access.2021.3071057
[26] Chen, Y., Wang, Y., Hu, F., Feng, L., Zhou, T., & Zheng, C. (2021). LDNNET: Towards Robust Classification of Lung Nodule and Cancer Using Lung Dense Neural Network. *IEEE Access*, *9*, 50301–50320. https://doi.org/10.1109/access.2021.3068896
[27] Petousis, P., Winter, A., Speier, W., Aberle, D. R., Hsu, W., & Bui, A. a. T. (2019). Using Sequential Decision Making to Improve Lung Cancer Screening Performance. *IEEE Access*, *7*, 119403–119419. https://doi.org/10.1109/access.2019.2935763
[28] Yu, H., Zhou, Z. E., & Wang, Q. (2020). Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model. *IEEE Access*, *8*, 86400–86410. https://doi.org/10.1109/access.2020.2992645
[29] Hussain, L., Aziz, W., Alshdadi, A. A., Nadeem, M. S. A., Khan, I. R., & Chaudhry, Q. (2019). Analyzing the Dynamics of Lung Cancer Imaging Data Using Refined Fuzzy Entropy Methods by Extracting Different Features. *IEEE Access*, *7*, 64704–64721. https://doi.org/10.1109/access.2019.2917303
[30] Gupta, S., Vundavilli, H., Osorio, R. S. A., Itoh, M., Mohsen, A., Datta, A., Mizuguchi, K., & Tripathi, L. P. (2022). Integrative Network Modeling Highlights the Crucial Roles of Rho-GDI Signaling Pathway in the Progression of non-Small Cell Lung Cancer. *IEEE Journal of Biomedical and Health Informatics*, *26*(9), 4785–4793. https://doi.org/10.1109/jbhi.2022.3190038