Approaches To Handle Imbalanced Dataset In Disease Predictive Classification: A Survey

Priyanka Pariyawala

(0009-0007-7836-3943)

Research Scholar,
Department of Computer Science
,Veer Narmad South Gujarat University, Surat, Gujarat,
India.
Email:- priyanka_pariyawala@yahoo.com,

Dr Pushpal Y Desai
Professor,
Department of ICT,
Veer Narmad South Gujarat University, Surat, Gujarat, India.
Email:- pydesai@vnsgu.ac.in

Abstract:

The integration of machine learning (ML) and artificial intelligence (AI) into medicine is reshaping disease prediction and diagnosis by enabling earlier detection and stronger clinical decision support. A key challenge, however, is class imbalance in medical datasets, where the scarcity of diseased cases compared to healthy ones leads to biased models and higher risks of misdiagnosis. This review explores imbalance-handling strategies ranging from traditional data-level approaches (undersampling, oversampling such as SMOTE and ADASYN) to algorithm-level solutions (cost-sensitive learning, ensemble methods) and hybrid frameworks. It also highlights recent advances with generative models (GANs, VAEs) and transfer learning that offer new opportunities for synthetic data generation and fairer predictions. By drawing insights across diverse applications—including cardiac, cancer, renal, and infectious diseases—the paper discusses both the promise and limitations of current methods, emphasizing the urgent need for reliable imbalance-handling solutions in medical AI.

I. INTRODUCTION

The connection between data science and healthcare brings revolutionary changes to medical disease diagnosis methods as well as management and forecasting. The utilization of machine learning (ML) and artificial intelligence (AI) technology makes it possible for medical professionals to establish predictive models based on enormous medical data to reinforce early diagnosis procedures and clinical treatment decision-making[1]. The dominant problem within this field involves unbalanced data distribution across datasets. Real-world clinical databases often present a disease pattern where patients with the condition are much fewer than those who do not have it which produces erroneous model predictions. The research investigates different methods to handle dataset imbalance particularly through an examination of their practical applications along with their constraints and their performance in disease predictive modeling[2].

A. Context of Predictive Disease Classification in Healthcare

Medical conditions are now predicted via predictive modeling based on patient information following the introduction of Artificial Intelligence (AI) and Machine Learning (ML) technology in healthcare[3].The identification of future illness development in patients through diagnostic elements including symptoms along with lab reports and imaging data and genetic markers makes up predictive disease classification. The method proves useful in detecting both long-term medical conditions and uncommon diseases and new infections. These models depend on excellent training data quality and proper data distribution for their clinical effectiveness and operational performance.

B. Prevalence of Class Imbalance in Medical Datasets

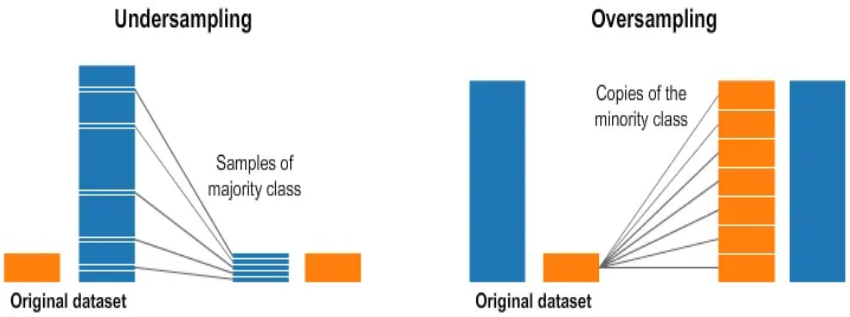
The healthcare domain faces widespread class imbalance because datasets normally contain many more records of healthy patients and common disease cases than those of rare and critical diseases. The datasets that focus on breast cancer contain numerous benign cases while having restricted numbers of malignant cases[4]. The data collection for confirmed infectious disease cases generally shows less abundance compared to negative or asymptomatic cases. The unbalanced distribution of data causes learning abnormalities because ML models do not properly identify traits from minority classes.

C. Impact of Imbalance on ML Model Performance

The performance of models deteriorates when they handle unbalanced datasets because the metrics for minority class evaluation such as recall and precision and F1-score suffer. Standard ML algorithms pursue overall accuracy optimization leading to learning models that show preference for the majority class. The high number of false negatives occurs when disease detection models incorrectly determine that no disease exists when it does. Errors of this nature present significant medical risks because of their dangerous nature[5]. The correct management of class imbalance stands as both a vital technical necessity and essential requirement for developing ethical and effective healthcare artificial intelligence systems.

II. NATURE AND IMPACT OF IMBALANCED DATA IN DISEASE PREDICTION

Medical and clinical datasets present the most severe form of class imbalance problems. The datasets display an unbalanced structure because they contain many more examples from the dominant group than from the less prevalent group. Machine learning models become less effective when dealing with class imbalance issues particularly when used for disease prediction and classification tasks.



A. Causes of Imbalance

Multiple reasons exist which result in the development of healthcare dataset imbalance.

- Rarity of Diseases:The occurrence rate of diseases like genetic disorders and rare cancers together with some infectious diseases remains naturally limited throughout the population[7]. The collection of sufficient representative data becomes challenging because of this condition.
- Low Sampling from Minority Classes:The clinical environment restricts minority-class sample collection because of practical and ethical restrictions. The medical staff reserves invasive tests for situations when their clinical assessment demonstrates their need.
- Data Collection Bias:Electronic Health Records (EHR) alongside other data sources originate from general hospitals and clinics which specialize in rare diseases which distorts the data distribution.
- Temporal or Demographic Skews:The inclusion of new diseases and age-specific or geographically limited diseases becomes limited in datasets that operate on broader scales.

B. Consequences of Imbalanced Data

The diagnostic quality of predictive models strongly suffers when data contains class imbalance[8]. The following adverse effects appear because of class imbalance:

- Biased Predictions:Most predictive algorithms optimize total accuracy through overall prediction which creates a fit that prioritizes the dominant class. The learning process fails to properly recognize minority class instances when this situation occurs.
- Misdiagnosis Risks:The detection of diseases faces significant risks from misdiagnosis when false negative results occur because such errors delay proper treatment and worsen patient health outcomes.
- Low Sensitivity and Specificity:Models developed from imbalanced data become unreliable during critical clinical situations because they demonstrate low sensitivity and specificity to the minority class.

- Poor Clinical Trust:Medical professionals have lower confidence in AI systems which repeatedly fail to detect uncommon severe medical issues unless these systems provide clear explanations for their diagnostic decisions.

C. Case Examples from Cardiac, Cancer, and Infectious Disease Datasets

The medical field faces specific difficulties because of class imbalance which produces performance distortions and unreliable models during critical healthcare operations[9]. The minority class which contains rare but important medical conditions receives limited representation during disease prediction tasks because this problem leads models to favor the majority class and overlook critical cases. A summary of representative data examples from cardiac, cancer and infectious disease fields appears in Table 1 which demonstrates imbalance characteristics and their effects on prediction accuracy and clinical results.

TABLE I. REFERENCE TABLE

Reference	Medical Domain	Dataset Characteristics	Nature of Imbalance	Impact on Prediction	Key Findings
Mustapha & Ozsahin (2024) [10]	Cardiac Diseases	Electronic health records for cardiovascular event prediction	Rare cardiac events (e.g., myocardial infarction) form minority	Models biased towards majority (healthy) class, resulting in low recall for critical events	Applying class imbalance handling methods significantly improved sensitivity without sacrificing overall accuracy
Chen et al. (2018) [11]	Cancer Diagnosis	Breast cancer datasets with predominance of benign cases	Malignant cases underrepresented compared to benign	Standard classifiers yield low detection rates for malignant tumors, risking delayed diagnosis	Use of advanced resampling techniques improved detection rates of malignant tumors substantially

Mulugeta et al. (2023) [12]	Renal Transplant Risk	Patient data from Ethiopian renal transplant recipients	Graft failure cases significantly fewer than successful grafts	ML models biased toward predicting graft success, missing early warnings of failure	Ensemble learning combined with imbalance handling improved prediction of graft failure risks
Wang et al. (2023) [13]	Chronic Pulmonary Disease	COPD patient data with skew towards mild or non-COPD cases	Severe COPD cases are minority	Class imbalance causes models to under-detect high-risk patients	Use of imbalance-aware algorithms enhanced early identification of severe COPD cases
Rodriguez-Almeida et al. (2022) [14]	Infectious Diseases	Small and imbalanced infectious disease datasets	Confirmed positive cases rare compared to negatives	Models overfit majority class; poor generalization on minority class	Synthetic data generation using GANs improved model robustness and minority class prediction

III. TAXONOMY OF TECHNIQUES TO HANDLE IMBALANCED DATA

Medical datasets containing unbalanced data distributions lead machine learning algorithms to develop bias that misidentifies crucial yet uncommon disease conditions. Various strategies exist to counter this issue which researchers group into data-level and algorithm-level approaches as well as hybrid solutions and recent deep generative models with transfer learning. The taxonomy system organizes a variety of approaches specifically designed to support predictive disease classification.

A. Overview of Technique Categories

The techniques for managing imbalanced datasets fall into five primary categories:

TABLE II. CLASSIFICATION IMBALANCE HANDLING APPROACHES[15]

Category	Objective
Data-Level Approaches	Modify training data to balance class distributions before learning begins
Algorithm-Level Approaches	Adapt the model training process to account for class imbalance explicitly
Hybrid Approaches	Combine data and algorithm-level strategies for synergistic improvements
Generative Models	Use deep learning-based generative techniques to synthesize realistic minority class data
Transfer Learning	Use pre-trained models or cross-domain knowledge to improve classification in imbalanced datasets

The benefits and barriers within each category depend on the combination of available data quantity and clinical framework in addition to the need for model interpretation.

B. Data-Level Approaches

The training data class distribution undergoes modifications through data-level methods to combat skewness. The methods used in data-level approaches are undersampling and oversampling and synthetic sample generation.

(a) Undersampling

The technique makes the majority class smaller to create a balanced data distribution.

- **Random Undersampling** selects random majority class samples for removal yet it discards important information in the process.
- **Tomek Links** identifies and removes borderline majority cases which are in close proximity to minority examples to create better decision boundary definitions.
- The **Edited Nearest Neighbours (ENN)** method removes data points that are difficult to classify through nearest neighbor measurements.
- The **Cluster Centroid** technique substitutes majority instances with cluster centroids to maintain data distribution in a condensed form.

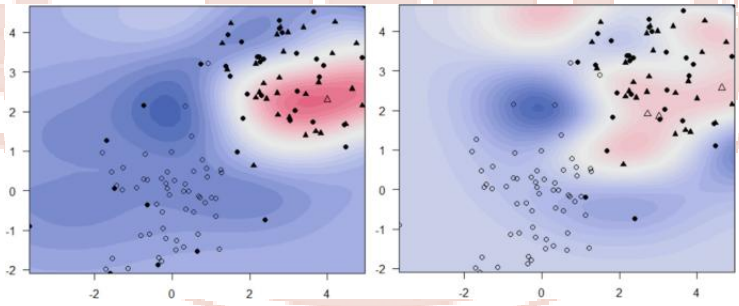
(b) Oversampling

Through this technique the number of minority class samples grows more numerous[16].

- The basic **random oversampling** method duplicates minority class instances yet creates the risk of overfitting the model.
- The **SMOTE (Synthetic Minority Oversampling Technique)** method develops artificial data points by connecting points from minority class groups.
- **Borderline-SMOTE** directs its oversampling efforts to the border areas where minority classes meet.
- The **ADASYN (Adaptive Synthetic Sampling)** method produces additional synthetic data for minority instances that are challenging to classify correctly.
- The **Safe-Level SMOTE** technique assigns safety levels to samples during neighbor synthesis to prevent noise contamination.

C. Algorithm-Level Approaches

The internal learning processes of machine learning models receive modifications through algorithm-level approaches to achieve effective class imbalance management. The most common approach in machine learning is cost-sensitive learning that gives higher weight to errors made while classifying minority samples. Healthcare practitioners widely apply this approach in medical imaging together with electronic health record (EHR)-based disease prediction because it maintains performance on vital yet uncommon outcomes.



Ensemble methods represent an essential category because they demonstrate exceptional ability to process imbalanced datasets. The combination of bagging with balanced bootstrapping in Balanced Random Forests trains individual base learners using more balanced subsets of the input data. The boosting algorithms including AdaBoost and XGBoost become more attentive to minority classes when they implement class weight integration. EasyEnsemble and BalanceCascade boundaries create several balanced data subsets which enables ensemble classifiers to find better generalization while preserving their ability to detect minority class patterns.

Support Vector Machines (SVMs) as well as Decision Trees and Neural Networks use class weight adjustment as a popular approach in machine learning algorithms. The algorithms achieve improved minority class detection because their penalty mechanism strengthens after misclassification occurs. Using cost-sensitive learning in cancer diagnosis and Alzheimer's disease prediction achieves superior results than conventional resampling methods that protect precision and recall.

D. Hybrid Approaches

The combination of data-level and algorithm-level approaches in hybrid strategies provides optimal imbalanced setting performance by achieving optimal classification results. The system effectively learns minority class patterns through Synthetic Minority Oversampling Technique (SMOTE) combined with cost-sensitive SVMs that maintains model generalizability. The model benefits from both strategies to process insufficient minority data while preserving its essential decision boundaries.

The most effective hybrid method in imbalanced data classification combines oversampling or undersampling methods with ensemble systems. The precise nature of these diagnostic systems becomes essential since false disease predictions can result in severe consequences thus making these techniques particularly effective for diagnosis systems. When ensemble models and balanced training subsets are used together they provide strong sensitivity performance[18].

The last approach to manage class imbalance in healthcare analytics is through pipeline-based models which implement systematic data management strategies. The first step involves applying SMOTE for data balancing then moving onto train GBMs or deep CNNs as complex models. Recent research in healthcare predictive analytics supports the increase in usage of these pipeline models because these allow deep learning architecture access while protecting the identification of crucial yet unusual patient classes.

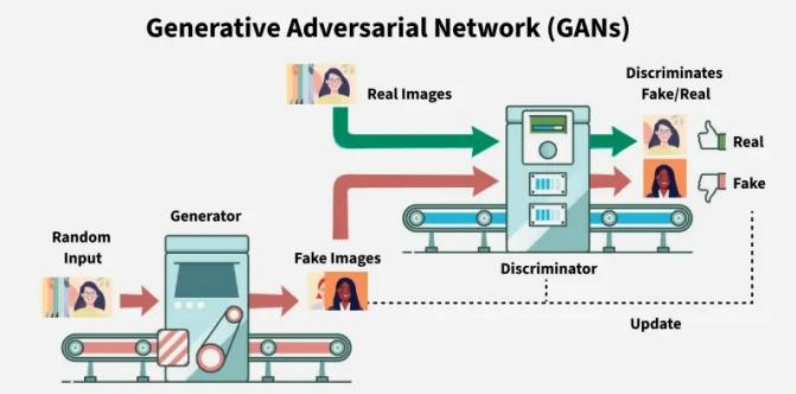
E. Generative Models and Transfer Learning

(a) Generative Models

Generative Adversarial Network (GAN):

Deep generative models serve as effective tools which generate high-quality synthetic data to address class imbalance problems in minority classes. GAN stands as one of the most widely used methods within this domain. Standard GANs prove effective at creating synthetic medical images together with structured electronic health record (EHR) data that serve to extend minority classes. The conditional GANs (cGANs) system builds upon

standard GANs by allowing the generation process to receive class labels which leads to the creation of relevant synthesized data. Specialized models BAGAN and SMOTifiedGAN use data augmentation together with imbalance correction features to enhance synthetic data quality.



Variational Autoencoders (VAEs):

The Variational Autoencoder (VAE) represents another essential generative method that learns condensed latent data representations from medical sources. VAEs have shown effective results for creating synthetic data from rare disease cases alongside filling gaps in clinical information databases. These models enable flexible generation of plausible patient data through their framework that preserves the initial minority class distribution in the original data. Generative models have demonstrated their effectiveness in medical applications like skin lesion classification and tumor segmentation through improved performance evaluation for rare minority cases.

(b) Transfer Learning

The application of transfer learning provides an effective solution to class imbalance through the use of pretrained models trained on extensive general-purpose databases. The models developed through ImageNet and MIMIC-III undergo specific fine-tuning to adapt to particular imbalanced medical application needs. The method transfers knowledge from extensive domains to specific underrepresented clinical areas without requiring extensive annotated datasets.

The concept of transfer learning produces beneficial outcomes in diagnosing early-stage lung cancer together with pediatric radiology and genetic disease classification applications. Modern domain adaptation methods establish connections between the pretraining source domain features and the fine-tuning target domain features to close the domain gap. The

techniques optimize representation learning to match target data better which leads to better predictive results particularly when dealing with minority classes in imbalanced datasets.

IV. TAXONOMY OF TECHNIQUES TO HANDLE IMBALANCED DATA

A. Undersampling Techniques

The majority class dataset gets reduced through undersampling to achieve a size that comes near the minority class count. This methodology reduces class bias while achieving better minority class sensitivity but may result in discarding vital predictive information that the model requires[20]. The selection of intelligent undersampling methods proves better than basic sample deletion strategies.

The imbalance ratio (IR) for dataset $D=D_{maj} \cup D_{min}$ calculates as the ratio of N_{maj} to N_{min}see (1):

$$IR = \frac{N_{maj}}{N_{min}} \gg 1 \qquad (1)$$

The objective of undersampling methods is to modify Nmaj until the imbalance ratio reaches IR=1 while maintaining important data points.

1. Tomek Links

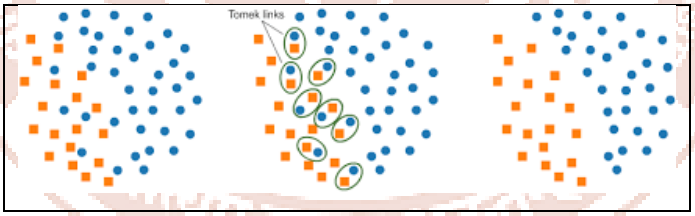


Fig. 4. Tomek Link for imbalance data[21]

A Tomek Link consists of two examples (xi,xj) when xi belongs to Dmaj and xj belongs to Dmin and satisfy as shown in (2):

$$\forall x_k \epsilon D, d(x_i, x_j) < d(x_i, x_k)$$
$$\text{and } d(x_i, x_j) < d(x_j, x_k) \qquad (2)$$

The distance metric d(·) most often uses Euclidean distances in this definition.

A Tomek Link develops when two samples from different classes identify each other as their closest neighbors which indicates they reside near the boundary decision. The removal of majority class instance xix_ixi from this pair helps decrease class overlap together with noise reduction.

Heart disease diagnosis benefits from Tomek Link analysis because it helps identify patients whose ECG signals resemble borderline cases between healthy and diseased patients. When borderline major class instances are removed it enables better detection of minority class characteristics including rare heart conditions such as arrhythmogenic right ventricular cardiomyopathy.

2. Edited Nearest Neighbour (ENN)

ENN removes the sample x∈Dx \in Dx∈D when the majority vote of its kkk-nearest neighbors produces a label different from its actual class[22].

Mathematically (3):

$$If\ y_x \neq mode(yNN_k(x), \text{then remove x} \qquad (3)$$

Where:

- Yx is the label of sample x
- The k-nearest neighbors of the sample x make up the set NNk(x).

ENN eliminates noisy and misclassified instances from the majority class when they deviate from local neighborhood patterns.

The prediction of Parkinson's Disease through healthcare data can be adversely affected by small changes in voice frequency or tremor data. ENN eliminates cases that stand outside typical boundaries and instances from the majority class which produce confusion during model understanding.

3. =Cluster Centroid Method

The unsupervised clustering technique (K-means is typical) performs on the majority class to create centroids which substitute each cluster group[23]. Formally:

Given:

- Majority class samples Dmaj
- Number of desired centroids k

Apply k-means clustering see (4):

$$\frac{min}{\mu 1,\ldots,\mu k} \sum_{i=1}^{k} \sum_{x \epsilon C_i} ||x - \mu_i||^2 \qquad (4)$$

Cluster Icontains all samples represented by Ci and its centroid value is μi.

The process involves substituting cluster groups of similar majority samples with their representative average (centroid) to maintain the distributional shape of the majority class.

The Healthcare Application utilizes gene expression datasets for cancer classification to maintain essential information from dense clusters of non-cancerous patients using centroid sampling methods.

TABLE III. COMPARATIVE OVERVIEW OF UNDERSAMPLING TECHNIQUES [24]

Technique	Mechanism	Mathematical Tool	Healthcare Use Case	Advantages	Limitations
Tomek Links	Removes borderline examples	Nearest-neighbor distance	Heart disease (ambiguous ECG signals)	Enhances decision boundaries	Ineffective on dense overlapping classes
Edited NN	Removes misclassified points	k-Nearest Neighbors	Parkinson's, diabetes detection	Noise reduction, better generalization	High computational cost
Cluster Centroid	Replaces groups with representative points	K-means clustering	Gene expression in cancer classification	Preserves distribution, reduces size efficiently	Risk of oversimplification

B. Oversampling Techniques

The distribution of medical data samples often becomes uneven when dealing with rare diseases or new condition stages that contain few instances of the minority class. Oversampling methods solve this issue by growing the minority class instances so the model can recognize its characteristics without developing biases toward the majority class. The following list presents major oversampling strategies which prove effective for healthcare AI research.

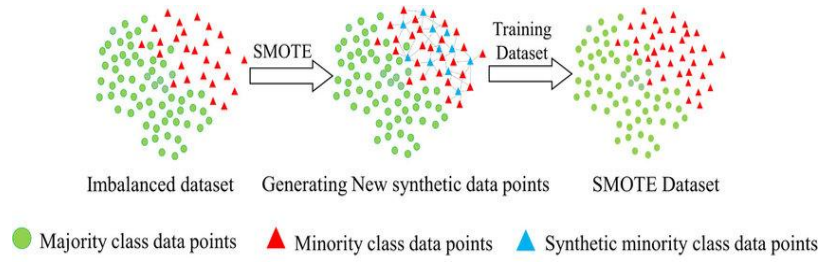


Fig. 5.Oversampling Techniques[25]

a) Random Oversampling

Random Oversampling stands as one of the first and easiest methods for managing class imbalance problems. The technique produces new minority class examples through random sample duplication until it reaches the required class balance. Random Oversampling remains straightforward to implement yet produces overfitting because the model views repeated minority class instances which diminishes its generalization ability.

The positive class minority problem has been addressed through this approach in healthcare disease detection systems such as diabetic retinopathy and pneumonia diagnosis. The approach yields better sensitivity but fails to provide novel information for the model which results in constrained performance improvements in complex dataset conditions.

b) SMOTE (Synthetic Minority Oversampling Technique)

SMOTE stands as one of the most popular advanced oversampling techniques which creates new synthetic minority class instances instead of making duplicates. SMOTE creates new plausible minority class examples through interpolating existing minority samples with their nearest neighbors inside their original class space. The model establishes better generalization abilities since it does not depend on memorizing actual data points.

The medical domain has implemented SMOTE successfully to classify breast cancer and predict liver disease and analyze cardiovascular risks. The recall and F1-score of minority classes improves when using SMOTE because this technique enhances dataset diversity thus reducing false negative outcomes that are vital in diagnostic models.

c) Borderline-SMOTE and ADASYN

Standard SMOTE encounters two major limitations because it produces synthetic samples both inside safe areas and noisy regions which led developers to create multiple enhanced versions.

Borderline-SMOTE generates new samples from minority class instances that exist near the decision boundary because these examples present the highest risk of misclassification. The model gains increased discrimination power in critical classification zones because synthetic samples are created exclusively

within these areas. In clinical early disease detection situations such as cancer and sepsis diagnosis the Borderline-SMOTE method proves beneficial because it targets challenging feature areas.

The Adaptive Synthetic Sampling method known as ADASYN allocates additional synthetic data creation to minority instances which demonstrate difficulty in learning because they exist

in areas dominated by majority class neighbors. The approach develops an active data-driven sampling technique which first targets hard-to-learn cases. Healthcare professionals employ ADASYN to predict sepsis onset while diagnosing rare diseases by concentrating model training on important yet scarce medical situations.

d) Safe-Level SMOTE

Safe-Level SMOTE implements risk-awareness through its ability to generate synthetic samples exclusively in dense minority-class areas which are less prone to overlapping with majority-class regions. The risk-awareness component in this variant prevents the generation of misleading samples that SMOTE typically produces in noisy or overlapping areas.

The tumor classification process using MRI or CT scans benefits from Safe-Level SMOTE because it produces synthetic instances exclusively in 'safe' areas which maintains minority class integrity and enhances class representation.

TABLE IV. COMPARATIVE INSIGHTS [26]

Technique	Key Idea	Application in Healthcare	Strengths	Weaknesses
Random Oversampling	Duplicate existing minority samples	Used in diabetic and pneumonia classification	Simple and fast	Overfitting risk
SMOTE	Create synthetic samples using interpolation	Common in cancer, liver disease, and heart disease	Improves generalization, reduces bias	May generate noisy or ambiguous samples
Borderline-SMOTE	Generate synthetic data near the decision boundary	Helpful in early disease detection	Focuses on critical classification regions	Ignores central (easier) minority samples
ADASYN	Adaptive focus on harder-to-learn samples	Effective in sepsis and rare condition prediction	Prioritizes challenging but important instances	Risk of amplifying noise
Safe-Level SMOTE	Generate only in safe, reliable minority class regions	Used in brain tumor detection via imaging	Avoids generating samples in noisy overlap regions	Needs safe-level parameter tuning

Oversampling stands as a crucial method to handle the class imbalance challenge that arises during disease predictive classification. Random Oversampling remains simple to execute yet the advanced techniques of SMOTE and its versions Borderline-SMOTE and ADASYN and Safe-Level SMOTE deliver sophisticated approaches for enhancing minority class learning. The selection process for oversampling methods depends on both dataset attributes along with medical application specifications. The implementation of oversampling requires special attention in sensitive medical fields including oncology, cardiology and infectious disease modeling to produce clinically valid and trustworthy prediction results.

V. DEEP GENERATIVE MODELS FOR DATA AUGMENTATION AND TRANSFER LEARNING

The increasing application of artificial intelligence in medicine shows a critical data imbalance issue that impacts disease predictive classification. Improvements in model fairness and sensitivity gained by means of conventional sampling methods such as SMOTE and ADASYN fail to hold original data distribution patterns in high-dimensional unstructured spaces that consist of medical images and time-series data. Deep generative models, specifically Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have been found to be useful as data augmentation and representation learning tools. These modeling techniques allow for the learning of complex healthcare data distributions and generate synthetic data samples that fill gaps in sparse classes effectively.

A. *Generative Adversarial Networks (GANs)*

The two neural networks composing GANs operate through the generator and discriminator which engage in a minimax game. The adversarial training process lets the generator create new data points which the discriminator must identify between real and synthetic instances. The adversarial training process enables the generator to create data instances which closely match the original dataset characteristics[27].

Standard GAN vs Conditional GAN (cGAN):

Standard GANs produce unstructured data without restrictions leading to ambiguous class outputs in their generated samples.

Both the generator and discriminator of Conditional GANs (cGANs) receive additional labels or conditions so they can generate data while observing specific classes. The capability of generating specific data types proves useful for healthcare applications when building well-balanced predictive models through targeted image creation (such as cancer-positive images).

Application in Disease Prediction:

The application of cGANs in medical imaging has generated synthetic images of rare conditions for brain tumors and diabetic retinopathy which expanded medical datasets to enhance classifier accuracy. The use of such settings produces enhanced minority class recall results alongside better F1-score metrics according to research findings.

a) Applications in Disease Image Synthesis

The tool has gained extensive popularity in medical imaging applications to produce realistic depictions of rare disease manifestations from small training sets. For example:

- GAN-generated chest X-rays serve as a tool for dataset balancing when detecting pneumonia.
- The use of cGANs for skin lesion image synthesis enables better melanoma classification models in dermatology practice.
- Artificial data from these samples enables models to become both precise and less sensitive to the dominant class.

b) Balanced GAN (BAGAN), SMOTifiedGAN

The problem of data imbalance has driven researchers to develop advanced GAN variants including:

- The Balanced GAN (BAGAN) system produces additional minority class examples for dataset balancing without altering the original data distribution. BAGAN generates samples that improve minority class detection rates without creating a biased model.
- The hybrid model SMOTifiedGAN merges GANs and SMOTE features to generate new samples although it retains the distribution of original data. The approach produces additional diverse and realistic samples which specifically benefit structured medical datasets including patient records and sensor information.

The two methods deliver effective results that boost classification precision when used to identify rare cardiac occurrences and detect epileptic seizures and screen for retinopathy conditions.

B. Variational Autoencoders (VAEs)

Probabilistic Variational Autoencoders function as two separate components to transform data into abstract latent vectors and afterward recover this information. VAEs learn from reconstruction loss along with regularization terms while GANs use adversarial loss for their learning process. The method produces smooth continuous latent spaces that allow generating new data instances through sampling[28].

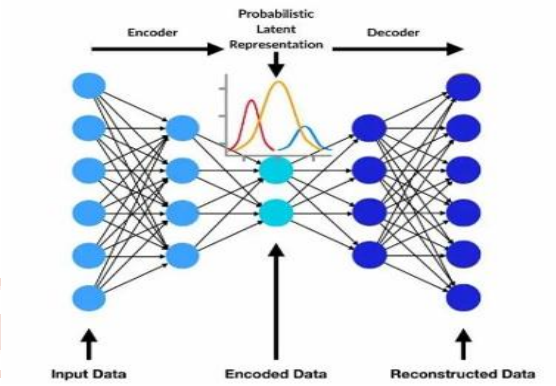


Fig. 6. Variational Autoencoders (VAEs) for imbalance data[29]

Learning Data Distribution:

VAE systems obtain true data distribution information through learning latent variable models. The interpretability and control capabilities of VAEs surpass those of GANs. VAEs demonstrate ideal functionality in situations where understanding data production plays a critical role because of their generative capabilities in synthetic data generation applications such as electronic health record (EHR) synthesis or gene expression profiling.

(a) Use in Healthcare Synthetic Data Generation

VAE technology works successfully in multiple healthcare field applications [30]:

- Researchers employed VAEs in EHR-based disease modeling to produce synthetic patient records of rare conditions such as Parkinson's and ALS because their scarcity stems from privacy restrictions and demographic constraints.
- VAEs create biological plausible gene expression profiles for cancer research subgroups that have limited representation in medical studies.
- VAEs generate new instances of time-series data including ICU monitoring or EEG signals which enhances model robustness for abnormal patient conditions.

VAEs enable healthcare applications to meet data privacy standards because they produce synthetic information which reduces the likelihood of patient identification.

VI. CONCLUSION

The problem of class imbalance throughout disease predictive classification affects both medical decision processes and patient wellness and AI system reliability in healthcare. Classic balancing methods of undersampling and oversampling are powerful tools for imbalance handling but create issues by causing information loss and overfitting problems. Cost-sensitive learning solutions and ensemble methods provide better performance though they require special calibration to perform best in different medical environments. The application of deep generative models including GANs and VAEs now presents new opportunities to create top-quality synthetic minority class samples while transfer learning enables the use of information from large datasets. Medical data imbalance remains a complex problem which no individual technique can solve completely. Researchers should direct their attention to developing context-aware solution combinations using various strategies which retain clinical readability alongside bias reduction. Data science progress must depend on strong partnerships between doctors, specialists, and data experts who will ensure both the technical success and ethical alignment and patient benefits for different patient demographics.

REFERENCES

1. P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” ACM Comput. Surv. (CSUR), vol. 49, no. 2, pp. 1–50, 2016.
2. K. M. Hasib et al., “A survey of methods for managing the classification and solution of data imbalance problem,” arXiv preprint arXiv:2012.11870, 2020.
3. P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, “Classification of imbalanced data: review of methods and applications,” in IOP Conf. Ser.: Mater. Sci. Eng., vol. 1099, no. 1, p. 012077, Mar. 2021.
4. L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, “Review of classification methods on unbalanced data sets,” IEEE Access, vol. 9, pp. 64606–64628, 2021.
5. G. A. O. G. D. Sambasivam and G. D. Opiyo, “A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks,” Egyptian Informatics J., vol. 22, no. 1, pp. 27–34, 2021.
6. M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” BMC Med. Inform. Decis. Mak., vol. 11, pp. 1–13, 2011.

7. T. Suresh, Z. Brijet, and T. D. Subha, “Imbalanced medical disease dataset classification using enhanced generative adversarial network,” Comput. Methods Biomech. Biomed. Engin., vol. 26, no. 14, pp. 1702–1718, 2023.
8. P. Yildirim, “Chronic kidney disease prediction on imbalanced data by multilayer perceptron,” in Proc. IEEE 41st Annu. Comput. Softw. Appl. Conf. (COMPSAC), vol. 2, pp. 193–198, Jul. 2017.
9. R. Blagus and L. Lusa, “Class prediction for high-dimensional class-imbalanced data,” BMC Bioinformatics, vol. 11, pp. 1–17, 2010.
10. M. T. Mustapha and D. U. Ozsahin, “Class imbalance and its impact on predictive models for binary classification of disease: a comparative analysis,” in Artif. Intell. Image Process. Med. Imaging, Academic Press, pp. 389–408, 2024.
11. Y. F. Chen et al., “Design of a clinical decision support system for fracture prediction using imbalanced dataset,” J. Healthc. Eng., vol. 2018, no. 1, p. 9621640, 2018.
12. G. Mulugeta, T. Zewotir, A. S. Tegegne, L. H. Juhar, and M. B. Muleta, “Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia,” BMC Med. Inform. Decis. Mak., vol. 23, no. 1, p. 98, 2023.
13. X. Wang et al., “Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data,” Comput. Methods Programs Biomed., vol. 230, p. 107340, 2023.
14. A. J. Rodriguez-Almeida et al., “Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets,” IEEE J. Biomed. Health Inform., vol. 27, no. 6, pp. 2670–2680, 2022.
15. A. Bhatia, A. Chug, and A. P. Singh, “Application of extreme learning machine in plant disease prediction for highly imbalanced dataset,” J. Stat. Manag. Syst., vol. 23, no. 6, pp. 1059–1068, 2020.
16. Y. Li, W. W. Hsu, and Alzheimer's Disease Neuroimaging Initiative, “A classification for complex imbalanced data in disease screening and early diagnosis,” Stat. Med., vol. 41, no. 19, pp. 3679–3695, 2022.
17. A. Gupta and S. Gupta, “Enhanced Classification of Imbalanced Medical Datasets using Hybrid Data-Level, Cost-Sensitive and Ensemble Methods,” Int. Res. J. Multidiscip. Technovation, vol. 6, no. 3, pp. 58–76, 2024.

18. M. M. Ahsan and Z. Siddique, “Machine learning-based heart disease diagnosis: A systematic literature review,” Artif. Intell. Med., vol. 128, p. 102289, 2022.
19. S. D. A. Bujang et al., “Imbalanced classification methods for student grade prediction: A systematic literature review,” IEEE Access, vol. 11, pp. 1970–1989, 2022.
20. X. Zheng, “SMOTE variants for imbalanced binary classification: heart disease prediction,” Univ. California, Los Angeles, 2020.
21. M. Talebi Moghaddam et al., “Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm,” BMC Med. Res. Methodol., vol. 24, no. 1, p. 220, 2024.
22. X. U. Duo and Z. X. U., “Machine learning applications in preventive healthcare: A systematic literature review on predictive analytics of disease comorbidity from multiple perspectives,” Artif. Intell. Med., 2024. [Online]. Available: https://doi.org/10.1016/j.artmed.2024.102950
23. M. Loey, G. Manogaran, and N. E. M. Khalifa, “A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images,” Neural Comput. Appl., pp. 1–13, 2020.
24. [24] S. Y. Yamaguchi, S. Kanai, A. Kumagai, D. Chijiwa, and H. Kashima, “Transfer learning with pre-trained conditional generative models,” Mach. Learn., vol. 114, no. 4, p. 96, 2025.
25. [25] S. Fayaz, S. Z. A. Shah, N. M. ud din, N. Gul, and A. Assad, “Advancements in Data Augmentation and Transfer Learning: A Comprehensive Survey to Address Data Scarcity Challenges,” Recent Adv. Comput. Sci. Commun., vol. 17, no. 8, pp. 14–35, 2024.
26. Y. Deng et al., “Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients,” NPJ Digit. Med., vol. 4, no. 1, p. 109, 2021.
27. G. Gonçalves, “A Comparative Study of Data Augmentation Techniques for Image Classification: Generative Models vs. Classical Transformations,” M.S. thesis, Univ. Aveiro, Portugal, 2020.
28. S. Chatterjee, D. Hazra, Y. C. Byun, and Y. W. Kim, “Enhancement of image classification using transfer learning and GAN-based synthetic data augmentation,” Mathematics, vol. 10, no. 9, p. 1541, 2022.

29. M. Majurski et al., “Cell image segmentation using generative adversarial networks, transfer learning, and augmentations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, pp. 0–0, 2019.
30. J. Su, X. Yu, X. Wang, Z. Wang, and G. Chao, “Enhanced transfer learning with data augmentation,” Eng. Appl. Artif. Intell., vol. 129, p. 107602, 2024.