IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 2, 2021, accepted March 26, 2021, date of publication April 6, 2021, date of current version April 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071389

Deep Generative Models to Counter Class Imbalance: A Model-Metric Mapping With Proportion Calibration Methodology

BEHROZ MIRZA[1], DANISH HAROON[1], BEHRAJ KHAN[1], ALI PADHANI[1], AND TAHIR Q. SYED[2]

[1]School of Computing, National University of Computer and Emerging Science, Karachi 75030, Pakistan
[2]Institute of Business Administration, Karachi 75270, Pakistan

Corresponding author: Tahir Q. Syed (tqsyed@iba.edu.pk)

ABSTRACT The most pervasive segment of techniques in managing class imbalance in machine learning are re-sampling-based methods. The emergence of deep generative models for augmenting the size of the under-represented class, prompts one to review the question of the suitability of the model chosen for data augmentation with the metric selected for the-goodness-of classification. This work defines this suitability by using newly-sampled data points from each generative model first to the degree of parity, and studying classification performance on a large set of metrics. We extend the investigation to different proportions of augmented data points for identifying the sensitivity of the metric to the degree of imbalance, leading to the discovery of an optimum proportion against the metric. The models used are GAN, VAE and RBM and the metrics include Precision, Recall, F1-Score, AUC, G-Mean and Balanced Accuracy. We offer a comparison of these models with the established class of data synthesizing counterparts on the aforementioned metrics. Deep generative models outperform the state-of-the-art on 5 metrics on multiple datasets and also comprehensively surpass the baselines. This work thereby recommends the following model-metric mappings: VAE for high Precision and F1-Score, RBM for high Recall and GAN for high AUC, G-Mean and Balanced Accuracy under various recommended proportions of the minority class.

INDEX TERMS Adversarial networks, anomaly detection, class imbalance, deep generative models, density estimation, generative variational auto encoders, instance hardness threshold, machine learning best practices, restricted Boltzmann machines.

I. INTRODUCTION

Class imbalance is a ubiquitous problem to machine learning tasks, where the class significant to a business or scientific need contributes a smaller proportion of the total data instances. Anomaly detection and its derivatives, namely fraud detection and money laundering, medical diagnosis, fault diagnosis, spam detection are major examples [20]. Extreme imbalance is common in financial fraud datasets, where the minority class may represent even fewer than 0.5% of the total instances [8]. To counter this, the most popular methods lie in the category of over- and under- sampling, which manage data volume so that classes are more equally represented, [29], increasing the classifier performance [23], [24], [25]. For the purpose of this document, the former is referred to as data-augmentative which either resample or generate data points, and the later as data-reductive which prune them away. Of the two, data-augmentative methods have gained widespread adoption among industry practitioners as these exploit the entirety of the information in the data. These can be broadly classified into 3 types, replicating (classic), *synthetic* (prevalent) and *generative* (novel). First one augments via replication, second method produces instances using locally-linear interpolation. The recently-introduced third category generate new instances by learning the data distribution. Instead of using an interpolation process these employ a wide range of algorithms, from Gibbs sampling and variational inference to game theory, [22]. This subtlety raises the question of how well to study the contribution or efficiency of the newly generated instances in terms of performance metrics.

It is well known that each metric gauges performance from a different perspective, which increases the significance of contextual metric selection. For instance it is

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

inappropriate to select *precision* which measures model exactness in lieu of *recall* which measures model completeness where a strong restriction on false negatives is required. This is to be followed by an optimal model selection and its mapping onto the chosen metric, but a general absence of guidance from open literature (an exception may be the overriding metrics definition for GAN performance assessment on high dimensional data, Section III-A) compels this decision to be on intuitive or preferential bases. For the purpose of this document, model selection refers to the type of generative model (Section III), and not the classifier, degree of complexity or hyper parameters. The high algorithmic variance in generative models motivates the search for a behavioural alignment of these with specific metrics and the following drawbacks in synthetic methods impede further exploration using the latter as a data-production base.

1) Synthetic instances may be not true data representatives as these lack a guarantee of instance novelty, randomization that may mimic the level of noise in data, [28].
2) These methods do not address the problem from probabilistic perspective. Therefore, obtained without a learned distribution; the synthesized examples lack interpretation and information required by a classification model, [4].
3) Any new point obtained by a linear synthesizer is created by local interpolation from neighbourhood. Localities in data are linear [72], therefore these points are not able to follow the curve of the data manifold.

The implication of imbalance on classification models have been well studied but that on performance metrics is relatively under- explored. Among the latter, works by [31] observe association between metrics with an emphasis on coherence of AUC metric. Works by [30] explore the effect on metrics performance as imbalance fluctuates via under-sampling on image datasets. However, the sensitivity of metrics with varying degree of sample proportionality using over-sampling (synthetic or generative) is yet to be explored. Further, a silence of the literature can be observed in proposing a unification of model-metric mapping with metric-sample proportionality.

This work therefore, fills in the research gap by proposing a quantifiable methodology, the Model-Metric Mapper (MMM), which presents a coherent and comprehensive prescription to the modeler in combating class imbalance. This includes: 1) model-metric mapping 2) calibrating metric-sample proportionality both, using 3) contextual metric selection.

A. CONTRIBUTIONS

The major contributions of MMM are:

1) It establishes an effective model-metric mapping, selecting the optimally performing models against the contextually relevant metrics.
2) It calibrates metric wise optimum data proportionality, exploring the degree of sensitivity of a metric to imbalance.
3) It advocates the use of deep generative models for data augmentation in the context of structured data.
4) It proposes a quantified methodology, guiding the modeler in their choice of data augmentation models.

B. PAPER ORGANIZATION

The paper is organized as follows: Section 2 reviews Prevalent approaches, Section 3 elaborates Deep, nonlinear models generative models, Section 4 discusses Performance evaluation metrics, Section 5 proposes the MMM methodology, Section 6 comprises Experimental setup, Section 7 tabulates and discusses Results, Section 8 sets MMM in motion and Section 9 concludes the paper.

II. PREVALENT APPROACHES

The problem of improving accuracy on skewed datasets was formulated in year 2000 at the first workshop held on the topic in American Association for AI conference, Japkowicz and Holte [34]. This section majorly discusses the prevalent augmentation and a significant subset of reduction methods. The next section is devoted to a comprehensive discussion on deep nonlinear models (generative methods), relatively new competitors to these methods. The methods are discussed as enlisted in below:

- SMOTE (SMT) - Methodical, linear, synthetic oversampling
- Instance hardness threshold (IHT) - Methodical, linear, under sampling
- Indiscriminate replication/removal - Random sampling
- Related works

A. SMOTE - METHODICAL, SYNTHETIC, LINEAR OVERSAMPLING

SMOTE - Synthetic minority over sampling, a linear technique developed by Chawla *et al.*, [41] and variants (Section 2.4) have been the prevalent oversampling choice by practitioners, Fernández *et al.*, [1]. The technique improves classifier generalization by creating new minority instances. This is in contrast with indiscriminate replication or undersampling techniques (Section 2.3, 2.4) where the former adds no new information and the latter removes it thus being detrimental to classification performance.

The Method SMOTE performs nearest neighbor identification followed by synthetic instance generation using Euclidean distance between feature vectors. A minority class instance (base) is randomly selected followed by its k nearest neighbours (support) identification. Each new instance is created as an additive operation on the feature space of the base instance with the product of an inter-feature difference between base-support pair and a random value. This interpolation creates synthetic instances along the line section between features.

The Strengths and Weaknesses: In contrast with indiscriminate oversampling which works in the data space, SMOTE work in the feature space. This makes decision boundaries flexible. However the linearity restricts these to low dimensional datasets. The technique fails to counter class overlap existing in multiple disjoint clusters; when coupled with undersampling it leads to significant information loss. The synthetic instances lack variance and are not equipped to cover curve manifolds.

B. INSTANCE HARDNESS THRESHOLD - METHODICAL, LINEAR UNDERSAMPLING

This under sampling technique by Smith *et al.*, [49] removes frequently mis-classified instances by computing their degree of hardness. The two forked approach first identifies recurrent mis-classified instances followed by unit level reason exploration. This singularity based dual analysis is in contrast to prevalent undersampling approaches (discussed later) operating at an aggregate level.

The Method: Equations 1 and 2 formulate the technique. Equation 1 calculates hardness for the instance $< x_i, y_i >$. Due to multiple learning algorithms used, the loss is a sum over these together with the weighing term $p(d|z)$. The term $p(y_i|x_i, d)$ defines the probability that d assigns the label y_i to x_i. Higher value means high probability prediction. The d is computed from $d = g(z, \beta)$ when learning algorithm g is executed with parameter β on z. Substituting it in Equation 2 yields $g_k(z, \beta)$. The probability $p(d|z)$ is estimated as $\frac{1}{|C|}$ while the probability $p(y_i|x_i, g_k(z, \beta))$ is a summation over multiple learning algorithms in set C.

$$H < x_i, y_i > = 1 - \sum_{H} p(y_i|x_i, d)p(d|z) \tag{1}$$

$$H_c < x_i, y_i > = 1 - \frac{1}{|C|} \sum_{k=1}^{|C|} p(y_i|x_i, g_k(z, \beta)) \tag{2}$$

IHT uses classifier-out-difference technique by Peterson and Martinez, [67] for measuring degree of predictive variance between classifiers followed by clustering. 20 algorithms are summarized into 9. An example is BayesNet, DecTable, Ripper, Simple Cart are summarized as Ripper. This forms the set C in equation 2 for computing instance wise hardness. Further, two entities per instance for feedback are computed: classifier score and classification frequency.

The Strengths and Weaknesses: IHT's distinction is instance mis-classification reason identification. Major reasons are class intersect, class tilt, and borderline complexity further subdivided into conflicting neighbours, disjunct size, class equilibrium, class probability and tree depths. The method also identifies a positive/negative instance-reason relation. Other undersampling approaches (Related works) rely on nearest neighbor variants which are vulnerable to class overlap. The major weaknesses of IHT are: may lead to information loss due to undersampling and heavy dependence on the classifier for computing hardness score.

C. INDISCRIMINATE REPLICATION/removal - RANDOM SAMPLING

The traditional approach to balance skewed datasets apart from the methodical counterparts discussed above is indiscriminate re-sampling. The dataset is balanced either via random replication of the minority instances or indiscriminate removal of the majority instances, Chawla, [50]. The former increases variance and the latter leads to over fitting, Pozzolo *et al.*, [27]. Further demerits include: in class overlap settings the former leads to meaningful information reduction and the latter leads to high mis-classification, Garcia *et al.*, [32], Cieslak and Chawla [33]. Estabrooks *et al.*, [51], have shown that several base classifiers report an improved accuracy on balanced datasets but linear separability being a prerequisite.

D. RELATED WORKS

Work by Douzas and Bacao, [7] propose Geometric-SMOTE (GMT), which produces instances in an ellipsoid surrounding the chosen minority instance. Work by Douzas *et al.*, [6] propose K-means-SMOTE (KMT) which combines clustering with SMOTE where the former identifies clusters and the latter generates samples in clusters leading to noise reduction. Reference [2] propose Adaptive-SMOTE which uses instance complexity to partition minority data before oversampling. The method gives better results than borderline extensions by Han *et al.*, [42] which generate instances near positive and negative neighbors. Work by Janbandhu *et al.*, [3] use Adasyn by He *et al.*, [45] for oversampling which creates different decision boundaries than SMOTE by producing samples near mis-classified instances. Douzas and Bacao, [39] use Self-organizing-map-oversampling, SOMO technique which preserves underlying manifold structure by creating two dimensional representation of the input space prior to applying SMOTE. Work by Koziarski *et al.*, [5] propose Radial-specific-over-sampling where the method identifies potential regions for minority instances creation. Work by Nekooeimehr and Lai-Yuen, [38] use Adaptive semi-supervised-weighted-oversampling, A-SUWO which uses cross validation for minority class cluster size identification and generates synthetic instances based on a weighting mechanism. Work by Bunkhumpornpat *et al.*, [40] use Density-based SMOTE, which uses DB-SCAN algorithm to discover clusters and generates instances along the shortest path from each minority class instance to a cluster's pseudo-centroid.

Undersampling approaches are discussed. Addabbo and Maglietta [26], propose parallel-selective-sampling which gives importance to majority instances near demarcation and eliminates those further away. Lin *et al.*, [21] use Clustering-balance to create majority class clusters equal to minority instances, than reducing the majority until it equals minority. Mani and Zhang, [46] use NearMiss and its variants which use nearest neighbour heuristics for undersampling. NearMiss-1 and NearMiss-2 select positive samples with smallest and farthest distance to negative samples respectively, while NearMiss-3 follows a two-step approach. Tomek-links removes the majority instance by identifying the disparate pair having the closest link. SMOTE-Tomek by Batista *et al.*, [53] and SMOTE-ENN by Batista *et al.*, [54] perform oversampling followed by undersampling, cleaning noisy instances.

III. DEEP, NON-LINEAR GENERATIVE MODELS

Deep learning based generative models have the capability of generating instances that have good likelihood guarantees with the parameters of the training distribution. The core idea behind generative modelling being; given a collection of high dimensional training instances, a model is able to do the following:

1) **Density approximation**: Given a large set of instances the model should be able to estimate the probability density function well enough to describe the data.
2) **New instance generation**: The model should keep the joint distribution of data over all variables, and have a random process that could generate new data instances from the estimated training distribution.

Overall, interest in deep generative models has spawned interesting outcomes namely synthetic music, art work and forged human faces. This study uses these models to generate instances for the minority class in an effort to combat class imbalance problem. The authors observe that these nonlinear models stand out from the traditional linear counterparts which were capable of producing new instances upper bounded by the variation present in the dataset. Three types of generative models are discussed below. Later the study provides results of the experiments performed on multiple imbalanced datasets using these models.

- Generative adversarial networks - (GAN)
- Variational autoencoders - (VAE)
- Restricted Boltzmann machines - (RBM)

A. GENERATIVE ADVERSARIAL NETWORKS

GANs by [70] follow a game theoretic approach where two models/players compete in an adversarial arrangement. The objective of the *generator* model is to generate instances analogous to training distribution while that of the *discriminator* model is to discriminate between actual and generated samples. Although being non-linear and generative, GAN differ from VAE and RBM as the latter adopt density approximation approach.

$$L(D_{dis}, G_{gen}) = min_{\Phi_g} max_{\Phi_d}[\mathbb{E}_{x\sim p_{data}} log D_{\Phi_d}(x) + \mathbb{E}_{z\sim p(z)} log(1 - D_{\Phi_d}(G_{\Phi_g}(z)))] \quad (3)$$

The Model: The equation shows the minmax objective of the bi-model network. The discriminator D with parameters Φ_d maximizes by making by $D(x)$ the actual sample as close to 1 and $D(G(z))$ the counterfeit as close to 0. However the generator G with parameters Φ_g minimizes by making $D(G(z))$ as close to 1. The purpose of the bi-model is to make the generator generate images which the discriminator presumes to be coming from the training distribution and not as counterfeit. The prevalent approach would have been to minimize the objective of the discriminator being correct but this leads to flat gradients where learning is required and vice versa. However a spin on the generator's objective leads to marked improvement where rather than minimizing the discriminator being correct; it is maximized to be incorrect.

The Strengths and Weaknesses: GANs derive their strength from game theoretic foundations with the bi-model competitive feature. The generated instances are high in quality having salient resemblance with the training data. Thus positioning the models as strong candidates in an image or transactions based generative setting. C-GAN by [35] for transaction oversampling, Be-GAN by [57] for crisp and high resolution, Convolutional-GAN by [59] for vector arithmetic based morphing, Cycle-GAN by [58] for reversible domain transfer, LS-GAN by [61] and Wasserstein-GAN by [60] are variances used for training stability. The major weaknesses of GAN are: these are difficult to train, lack quantified performance assessment, does not use density approximation and follow a complicated inversion mechanism.

Overriding metrics for performance assessment Works by [17] and [11] propose a quantified performance assessment of GAN by expressing precision and recall differently. These consider FID metric [19] as uninformative due to its qualitative nature. Works in [17] use density modelling with mode change while works in [11] later arguing this as ambiguous articulate non-parametric manifold with truncation for metric estimation and quality/variation tradeoff, [12], [13]. Designed for assessing GAN performance on high dimensional data, both approaches use balance datasets. The former does introduce imbalance via mode change (class addition/removal), while the latter is silent on the subject.

B. VARIATIONAL AUTO ENCODERS

VAE by [69] are built on the idea of fusing autoencoders with probabilistic graphical models. The objective is to estimate and encode an intractable probability density via an understandable surrogate density, e.g. a Gaussian, then minimize the Kullback-Leibler divergence. These models are different from traditional autoencoders as they induce probability thus shifting the paradigm from deterministic to a random.

$$p\beta(x \mid z)] = \frac{p\beta(x \mid z)]p(z)\beta}{p\beta(x)} \quad (4)$$

Computing the posterior $p\beta(x \mid z)]$ in graphical models has been intractable and has been estimated using Gibbs sampling or variational inference. VAE use the latter which works on maximizing the lower bound.

$$L = E_{q(z)}[logp\beta(x \mid z)] - KL[q\phi(z \mid x) \parallel p\beta(z))] \quad (5)$$

The model The encoder and decoder using parameters ϕ and β respectively, produce distribution parameters μ and Σ. These are used to sample latent factor representation $(x \mid z)$ and reconstruction $(z \mid x)$ from the encoder and decoder respectively. These being differentiable lead to maximizing the lower bound. Equation 5 shows the loss function with the first term being reconstruction error and the second being the KL divergence regularizer together making up the lower bound. The encoder assumes a tractable Gaussian $q\phi(z \mid x)$

using KL divergence to make this close to $p\beta(z \mid x)$. The objective is to make prior closed to the posterior. This leads to deriving the second term in the loss function. The first term is expectation maximization of the conditional distribution $q\phi(z \mid x)$ with respect to q(z). As z being Gaussian makes the decoder a minimizer of reconstruction error.

The Strengths and weaknesses: The stochastic characteristic distinguishes VAE encoders from its traditional counterparts. The latent encoding z being sampled from Gaussian μ and Σ parameters transform the disjointed representation into a continuous one. This paves way for a generative model to not only replicate but generate interesting image variations. The combined optimization of the two terms namely the KL divergence and the reconstruction loss induces a two-fold effect. The regularizer enforces yet random but densely packed encodings and the loss encourages clustering of similar encodings. This leads to decoder generating instances having local variation within similar samples and interpolating feature mixes between dissimilar clusters. The major weaknesses of VAE are: these generate blurry outputs at times, have subprime variational issues due to amor-tization and approx-imation gaps and produce gradients having high variance.

C. RESTRICTED BOLTZMANN MACHINES

These are unsupervised generative models proposed by [68], designed as a symmetrical arrangement of binary stochastic neurons where two layers form a bipartite graph using non-linearity. Though deep models have produce sound generalization results but training and parameter optimization has been a challenge. Initialization with large weight values leads to poor local minima problem, while small weights leads to small gradients. But, with calibrated weights, learning algorithm performs well. This does require learning one layer of features at a time and captures strong high-order correlations of units in the layer below.

$$E(v,h) = -\sum_{i\epsilon pixels} b_i v_i - \sum_{j\epsilon features} b_j v_j - \sum_{i,j} v_j h_j w_{ij} \quad (6)$$

$$p(v,h) = \frac{exp(-E(v,h))}{Z} \quad (7)$$

The Model: RBM defines distribution over visible unit v with latent variables h via energy function E. As shown in equation 6, negative w leads to high energy with a decrease in probability and vice versa. Energy and probability being reciprocal. The function gives the probability distribution $p(v, h)$ shown in Equation 7. The challenge being: the partition function Z is the sum over all values of v and h. These are binary, so Z can take many values leading to an exponential sum over the numerator, making it intractable. To counter this [68] proposed *contrastive divergence*. The technique uses Gibbs sampling to approximate joint distribution when direct sampling is difficult. Alternating between layers, given one unit in visible layer, all units are independent in hidden layer, values in one layer be sampled given a value in another layer.

The Strengths and weaknesses: RBM are highly expressive models equipped with the capacity to encode a distribution without compromising computational efficiency. Symmetric connectivity between visible and hidden units makes faster algorithms likely. Unsupervised pre-training moderates parameter values in suitable ranges which makes back propagation efficient. Layer wise stacked unit creates deep belief networks which serve as meaningful feature extractors. The weaknesses of RBM are: these are tricky to train, vulnerable to local minima trap, use partition function which is difficult to approximate.

D. RELATED WORKS

Works by Engelmann and Lessmann *et al.*, [4] use conditional-WGAN, a type of GAN on multiple credit-scoring datasets for minority class generation. Fiore *et al.*, [8] use GAN for generating minority instances on financial anomalies dataset and outperforms traditional SMOTE. Zheng *et al.*, [9] combine GAN with adversarial denoising autoencoder for countering imbalance in telecom fraud setting. The model outperforms state-of-the-art namely bayesian belief network, fuzzy inference and deep auto-encoder models. Douzas and Bacao, [35] follow generative oversampling using conditional-GAN on real and synthetic datasets. Park *et al.*, [18] propose tab-ular-GAN which using a common structure for tabular and image data converts tabular rows into 2D matrix for convolution.

Tingfei *et al.*, [14] use Variational auto encoders for over sampling minority class and outperform synthetic models on financial datasets. Islam *et al.*, [15] use VAE for generating accident events, in highly imbalance setting and compared it with multiple Smote extensions. Dai *et al.*, [16] use contrastive variant of Variational auto encoder for generating under represented class on clinical datasets surpassing linear models. Works by Guo *et al.*, [10] use Gaussian Mixture VAE on high dimensional time-series data for generating minority class.

Works by Zieba and Tomczak [37] use Restricted Boltzmann Machine in an imbalance credit rating evaluation mechanism. Boltzmann encoded adversarial machines by [64] extend RBM where the model is trained against an adversary making it capable to discriminate between training and generated instances. Works by [65] use Gaussian-Bernoulli models as an extension to RBM. These are equipped with processing continuous data with improved gradients and used for oversampling.

IV. PERFORMANCE EVALUATION METRICS

The metric frequently used for evaluating model performance is accuracy. Being good at summarising, it is uninformative on imbalance data. As it weighs confusion matrix quadrants equally by measuring fraction of correct to total predictions thus does not provide an adequate measure on performance of minority instances. The following discussion highlights metrics preferred over accuracy in this context. However, due to different formulation and model gauging perspectives of each, the importance of contextual relevance of the metrics is also discussed.

1) *Precision*: High precision reports *exactness* of the model using positive predicted rate with lower number of false positives. The metric is highly useful where lowering false alerts is mandated. An ex: Spam detection in emails require model with high precision, otherwise incorrect classification of non-spams(true negatives) as spams(false positives) will result in valuable information loss.
2) *Recall*: High recall reports *completeness* of the model using true positive rate. The metric is useful where the objective is to capture majority of the minority instances with minimum false negatives. An ex: Infectious disease or fraud/money laundering detection requires model with high recall, where a weak one may predict infected patient/ fraudulent transaction (true positives) as healthy/genuine (false negatives), having devastating consequences.
3) *F1-score*: High F1-score signifies *moderateness* of the model, balancing precision and recall using harmonic mean of the two. The metric is useful where a moderation between false positives and false negatives is required. An ex: Models in information retrieval applications require high F1-Score where a successful search has to maintain the balance of including relevant and not including irrelevant documents.
4) *AUC*: Significance of AUC is its sensitivity towards model *rank ordering*. Higher scores highlight quality of demarcation. The metric is useful where grading predictions are more significant than producing probabilities. An ex: in terminal disease prediction setting thresholds is tricky, where a conservative/high value may lead to skipping actual patients and vice versa. Here AUC may be the preferred choice.
5) *Balanced accuracy*: High balanced accuracy reports model *comprehensiveness*. The metric uses true negatives together with true positives. Consideration of both minority and majority class makes models with high balanced accuracy relevant where completeness is mandated. It is also used where the test set is single or imbalanced. An ex: A fault detection model for newly manufactured machine, will use balanced accuracy.
6) *G-Mean*: The metric reports model *egalitarianism* or *homogeneity* as it boost accuracy on each class while maintaining balance among the accuracies. The metric is root of the product of two attributes, sensitivity and specificity. It is used where less conservative measure than harmonic mean is required. An ex: Models for stock market growth or investment portfolio yield predictions where the trends are non-linear or proportional use G-Mean.

Therefore it can be summarised, contextual metric selection is required. This will ensure the attribute being measured; is the one required to assess model's performance.

V. THE MMM METHODOLOGY

The proposed *Model-Metric Mapper methodology (MMM)* is conceived on the idea that an appropriate model selection is required following a suitable metric identification. The methodology adds on that metric specific minority-to-majority proportion is further required to get an optimum performance. The methodology is introduced below with its distinguishing features followed by the artifacts.

1) *Distinguishing features*: MMM exhibits 2 ground breaking features highly significant to data imbalance. These include:
   - *Model to metric mapping:* Designed to work with heavily imbalance datasets, the MMM guides a practitioner in relevant model selection for sample generation based on the required metrics. This work views that architecturally and algorithmically dissimilar models behave differently on distinct metrics (empirical evidence in Section VII). This is significant in imbalance context where performance on a specific metric is a key determinant in model selection. Hence, an informed model selection is required. To elaborate, selecting a model with high precision in an infectious disease environment may be useless where a high recall one would have been the obvious choice. Thus, MMM transforms the prevalent model selection approach from an intuitive/preference based to an informed one.
   - *Metric wise sample proportionality calibration:* Together, with the appropriate model selection; MMM identifies the optimum minority-to-majority proportion against specific metric using quadrant wise calibration search. This work views that metrics are sensitive to sample proportionality, hence a metric specific proportion is to be identified. (empirical evidence in Section VII and discussed in Section VIII). This being significant as prevalent approaches of increasing the minority or decreasing the majority for mere balancing and without considering the required metric wise proportionality may not yield optimum results.
2) *The artifacts*: MMM is constituted as two independent modules or artifacts. The first artifact systematically generates minority instances from three different deep generative models. The second subsumes the first while adding a systematic reduction of the majority class. The artifacts are discussed:
   Artifact I - Generative calibrations is designed on the *generative* concept with the premise that instances generated from deep generative models are grander than the competing synthetic or re-sampled counterparts in the context of being novel and emitted from learnt joint distributions. This brings these in close proximity to the original ones. The artifact using quadrant based calibration employ the generated instances in search of an optimum metric wise majority to minority proportion and the model that yields it.

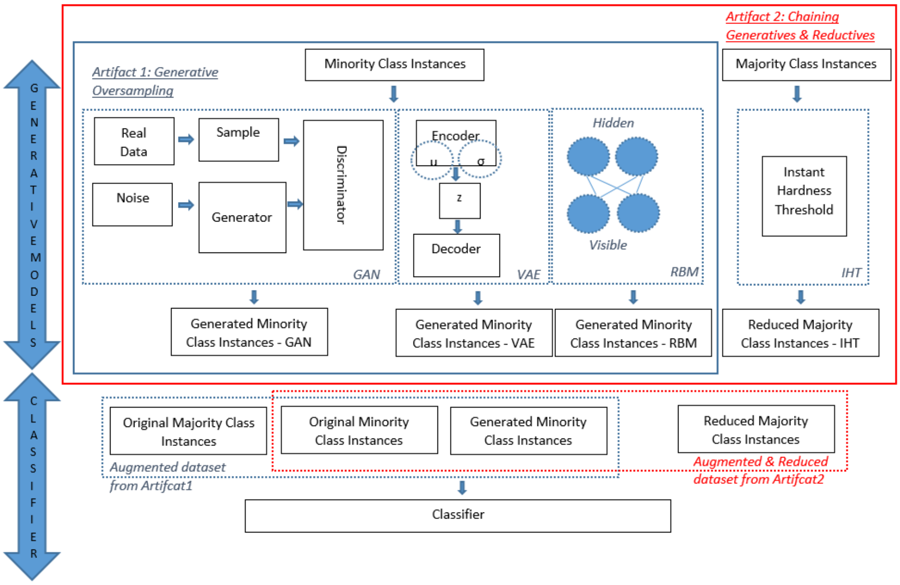

Artifact 2: Chaining Generatives & Reductives
Artifact 1: Generative Oversampling
Minority Class Instances
Majority Class Instances
GENERATIVE MODELS
Real Data
Sample
Noise
Generator
Discriminator
GAN
Encoder
μ
σ
z
Decoder
VAE
Hidden
Visible
RBM
Instant Hardness Threshold
IHT
Generated Minority Class Instances - GAN
Generated Minority Class Instances - VAE
Generated Minority Class Instances - RBM
Reduced Majority Class Instances - IHT
CLASSIFIER
Original Majority Class Instances
Original Minority Class Instances
Generated Minority Class Instances
Reduced Majority Class Instances
Augmented dataset from Artifcat1
Augmented & Reduced dataset from Artifcat2
Classifier

FIGURE 1. The MMM methodology with orchestration for deep generative & reductive methods.

The artifact has three streams each encompassing an architecturally and algorithmically different generative model namely GAN, VAE and RBM. VAE and RBM are density estimators but as density function being intractable the former use variational inference and the latter Gibbs sampling for approximation. GAN adopt a game theoretic approach rather than working with specific density function which makes it different from the two models.

$$\chi = \alpha(|\chi_{min}| + |\chi_{min\_gen}|) + 1 - \alpha(|\chi_{maj}|) \tag{8}$$

Equation 8 represents Artifact I. χ_{min} are the original and χ_{min_gen} the generated minority instances respectively. These being generated as distinct sets from the 3 deep models. χ_{maj} are original majority instances. α is the *coefficient of proportionality* with values set as $\{1/4, 1/2, 3/4$ and $4/4$ or $1/1\}$. This is used for governing the minority class instance generation proportion. The artifact uses quadrant wise calibration search by combining the majority and minority instances using the coefficient of proportionality and measuring against the mentioned 6 performance metrics. The search cycle continues until an optimum metric wise proportion and model is identified.

Artifact II - Chaining generative and reductive calibrations Is designed on the *generative+reductive* concept with the premise that coupling minority instances generation with majority instances reduction in a systematic order may have the following effects. First, the reduced noise and induced novelty may lead to a different model-metric mapping than Artifact I. Second, majority to minority proportions may vary against the ones identified in previous Artifact. Therefore, Artifact II connects or chains IHT, the undersampling technique (already discussed) with contemporary instance generation from Artifact I. This constitutes the two links namely generatives and reductives in the artifact's chain.

$$\chi = \alpha(|\chi_{min}| + |\chi_{min\_gen}|) + 1 - \alpha(|\chi_{maj\_iht}|) \tag{9}$$

Equation 9 represents Artifact II. χ_{maj_iht} are the majority and χ_{min_gen} minority set, ensuing from reduction and generation models respectively. χ_{min} are the original minority instances. α is the *coefficient of proportionality* with values determined as $3/22 \approx 1/7$, $6/19 \approx 1/3$, $9/16 \approx 1/22$ and $12/13 \approx 1$. This is used for governing the minority class instance generation proportion. The artifact uses quadrant wise calibration search by combining reduced majority with the original and generated minority instances using the coefficient of proportionality and measuring it against 6 performance metrics. The search cycle continues until the optimum metric wise proportion and the model is identified. As for IHT, when applied to majority class, it assigns and removes instances of lower probabilities using factors namely class skew, overlap and decision boundary complexity. Artifact II differs from Artifact I as the later focuses on instance generation while the former couples reduction with generation. This is in addition to the fact that unlike the later, the former alters both the minority and majority instance counts.

Therefore, MMM recommends metric specific informed model selection with calibrated sample proportion, localizing the model and required data proportion to the metric level. The methodology strengthens its recommendation using 2 independent artifacts both leading to similar conclusions.

TABLE 1. Datasets.

Name	Access	Original Dataset			
		Maj.	Min.	Total	Min%
CC	Public	284,315	492	284,807	0.17
GMC	Public	139,974	10,026	150,000	6.68
PH	Public	115,304	1,296	116,600	1.11
SS	Public	194,198	50,858	245,057	20.7
AM	Prop.	575,000	5,900	580,900	1.01

VI. EXPERIMENTAL SETUP

Experimental Setup The sub-section comprises of dataset details, abbreviations, model parameters, evaluation metrics, and experimental nomenclature. Table 1 enlists the 4 datasets used later in the experiments.

Datasets As shown in table 1, experiments are conducted on 4 public and 1 proprietary dataset. *Credit-card-fraud-detection* dataset was collected by ULB (Université Libre de Bruxelles), [2]. It comprises of genuine and fraudulent transactions by European cardholders. It is highly imbalanced with 0.18% frauds. *Give-me-some-credit* dataset classifies risky financial borrowers, [73]. Minority accounts for 6.68% making it highly imbalance. *Protein-homo* dataset categorizes protein sequence comparability, [74]. Being highly imbalance, the minority class is 1.11% of the total. *Skin-no-skin* dataset covers skin segmentation task, [75]. Anomaly being 20% makes it imbalance. *Anti-money-laundering-cases* is a proprietary dataset comprising of financial transactions flagged as cleared and laundered. Being heavily imbalance as the minority class constitutes 1.01% of the total volume. All the datasets are included keeping in view volume, imbalance and tabular structure.

Abbreviations The work uses following abbreviations against the models/datasets. Models: Generative adversarial networks (GAN), Variational auto encoders (VAE), Restricted Boltzmann machines (RBM), Smote (SMT), KMeans-Smote (KMT), Geometric-Smote (GMT), Instance hardness threshold (IHT). Models with IHT: (GAN-i), (VAE-i),(RBM-i),(SMT-i),(KMT-i),(GMT-i). Datasets: Credit-Card (CC), Give-me-some-credit (GMC), Protein-Homo (PH), Skin-No-Skin (SS) and Anti-money-laundering-cases (AM). *Model-wise parameters* used in the experiments are enlisted in Table 2.

The effect of the use of different classifiers is made invariant by the use of a single classifier, the-industry-standard, XGBoost across all experiments. The parameters used are depth = 5, weak learners = 100 and learning rate = 0.1.The default implementations of the models are used as provided in GAN [78], VAE [76], RBM, [77], SMT [41], KMT [6], GMT [7], and IHT at [49].

Evaluation metrics based on highly imbalance nature of datasets, the evaluation metrics used are Precision, Recall, F1-score, AUC, G-Mean and Balanced accuracy.

TABLE 2. Model configurations.

	GAN	VAE	RBM
Model type	Wasserstein	Kingma	Gaussian Bernoulli
Loss function	MinMax Cross entropy	Negative log likelihood with (KL)	Mean squared
Optimizer	Adam		
Epochs	1000		
Batch size	64		
Learning rate	0.0001		
Momentum	0.5		

Experimental Nomenclature The experiments are divided into 3 sets namely

- Experiment set-B 'Baseline comparison': compares models from Artifact I and II with baselines. The results are shown and discussed in Section 7.1.
- Experiment set-I 'Generatives vs Synthetics': compares state-of-the-art synthetic with Artifact I's generative models using synthetic and generative oversampling respectively. These models collectively fall into 'Data Augmentative Category' with results shown in Section 7.2.
- Experiment set-II 'Generatives + Reductives vs Synthetics + Reductives': compares state-of-the-art synthetic with Artifact II's generative models both employing a common undersampling technique. These models collectively fall into 'Data Augmentative + Reductive Category' with results shown in Section 7.3.
- An inter-category comparison of leading model from each of the two categories is performed to identify the overall top performer against each metric. This is discussed in Section 8.

VII. RESULTS

The results section is divided into 4 segments:

1) Baseline comparison
2) Generatives vs synthetics
3) Generatives+reductives vs Synthetics+reductives
4) Training efficiency

A. BASELINE COMPARISON

The baseline comprises of the dataset in its nascent form and is compared with both generative and generative+reductive approaches from Artifact I and Artifact II respectively. For comparison the Artifacts produce training data with 1:1 minority-to-majority ratio.

The objectives being:

1) To supports and rationalize the argument that balancing leads to an increase in performance metrics in general. Artifact I generates the minority while Artifact II reduces the majority together with increasing the minority to achieve training data balance.

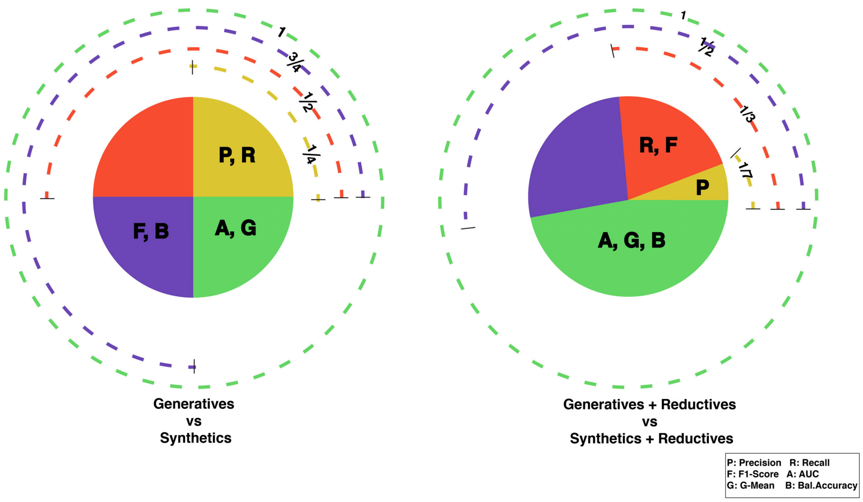


FIGURE 2. **Calibrating metric-sample proportionality.**

2) To evaluate the behaviour on 6 different performance metrics. This being significant as these metrics are preferred over traditional accuracy metrics for reporting in imbalance settings.

The comparison with baselines is performed on 6 distinct metrics over multiple datasets for generalizability. As shown in Table 3, approaches from both Artifacts surpass the baseline comprehensively over 6 metrics namely Precision, Recall, F1-Score, AUC, G-Mean and Balanced accuracy. Generative oversampling with maximum % increase of 1.71, 452.9, 7.46, 15.2, 19.7, 13.9 while chaining generatives and reductives with maximum % increase of 1.16, 435.29, 51.67, 30.55, 82.39, 30.55 beats the baseline. Thus providing a strong case that balancing either by generation or generation+reduction leads to an increase in performance.

This work further identifies that specific metric may mandate a more precise minority-to-majority ratio than mere balancing for further improvement as will be shown in next subsections.

B. GENERATIVES VS SYNTHETICS

This section compares generatives from Artifact I with state-of-the-art synthetic models including SMT and its current extensions KMT and GMT. The comparison is performed proportion wise. The objectives of this comparison being:

1) Compare 2 data augmentative techniques using 6 metrics on multiple datasets. To empirically identify the benefit of using generative models against synthetic counterparts.
2) Recommend a model-metric mapper. This works argues that model performance varies per performance metric. Therefore, an effective model should be selected based on the given metric.
3) Search for an optimum minority-to-majority ratio. The authors of this work view that identifying and maintaining precise metric specific proportionality together with the yielding model improves performance.

The results are tabulated in Tables 4 to 9 and proportionality summarized in Figure 2. *Top, second best scores and proportionality quadrants use the same colours for analogy*. Table 4 shows results on Precision metric. On all 5 datasets, VAE leads with scores of (0.87,0.54,0.89,0.9,0.88) against the synthetic KMT with (0.83,0.53,0.88,0.89,0.84). The best scores are found where minority-to-majority ratio is 1/4 or the 1st quadrant. VAE reports a maximum increase of 4%. On Recall, RBM clearly leads on all datasets with scores (0.95,0.97,0.96,0.98,0.96) as shown in Table 5. SMT follows with (0.88,0.54,0.9,0.96,0.86). Majority of the top scores are reported in the 1st quadrant with proportionality 1/4. RBM reports a maximum increase of 43%. Table 6 shows results on F1-Score. VAE surpasses on 4 datasets with (0.78,0.82,0.94,0.84) while SMT leads on 1 with (0.94). The 2nd best scores are (0.75,0.31,0.8,0.92,0.8) by GAN,SMT,KMT,SMT and KMT respectively. Majority of the best results are found where minority-to-majority ratio is 3/4 or 3rd quadrant. VAE reports a maximum increase of 5%. Table 7 shows the results on AUC metric. GAN excels on 3 and SMT on 2 datasets with scores (0.94,0.59,0.88,0.97,0.93) and (0.93,0.72,0.94,0.95,0.92) respectively with VAE following closely. The prime results are found in the 4th quadrant where minority equals majority. The maximum increase reported by GAN is 2%. Table 8 shows results on G-Mean metric. GAN and VAE closely follow on all the datasets. The best results are found in the 4th quadrant with proportionality 1/1. The Balanced accuracy metric results are reported in Table 9. GAN leads on 3 and SMT on 2 datasets with scores (0.93,0.61,0.87,0.98,0.93) and (0.91,0.72,0.94,0.92) respectively, with VAE as runner up. The top scores are

TABLE 3. Comparing baselines.

		Generatives				Generatives+Reductives			
	Precision								
	Base	VAE	GAN	RBM	%Δ	VAE-i	GAN-i	RBM-i	%Δ
CC	0.85	0.85	0.58	0.02	⟷	0.67	0.54	0.84	↓ 21.51
GM	0.51	0.51	0.52	0.12	↑ 1.71	0.27	0.27	0.12	↓ 48.11
PH	0.86	0.86	0.75	0.18	⟷	0.54	0.43	0.18	↓ 37.24
SK	0.9	0.9	0.75	0.31	⟷	0.91	0.79	0.2	↑ 1.16
AM	0.80	0.80	0.78	0.09	⟷	0.54	0.43	0.21	↓ 32.5
	Recall								
CC	0.63	0.7	0.81	0.94	↑ 49	0.8	0.82	0.65	↑ 30.1
GM	0.17	0.19	0.18	0.94	↑ 452.9	0.64	0.64	0.91	↑ 435.29
PH	0.73	0.75	0.75	0.85	↑ 16.4	0.82	0.82	0.91	↑ 24.65
SK	0.9	0.88	0.98	0.92	↑ 8.8	0.95	0.97	0.97	↑ 7.77
AM	0.85	0.73	0.73	0.94	↑ 10.58	0.81	0.81	0.89	↑ 4.71
	F1-Score								
CC	0.73	0.74	0.66	0.01	↑ 2.32	0.73	0.73	0.21	⟷
GM	0.25	0.27	0.26	0.01	↑ 7.46	0.39	0.39	0.11	↑ 51.67
PH	0.79	0.79	0.75	0.14	⟷	0.67	0.58	0.15	↓ 15.48
SK	0.9	0.89	0.85	0.23	↓ 1	0.9	0.87	0.17	⟷
AM	0.77	0.77	0.77	0.06	⟷	0.68	0.55	0.21	↓ 13.25
	AUC								
CC	0.82	0.85	0.94	0.68	↑ 15.2	0.9	0.91	0.55	↑ 11.19
GM	0.58	0.59	0.58	0.59	↑ 2.05	0.76	0.76	0.5	↑ 30.55
PH	0.86	0.88	0.88	0.8	↑ 2.22	0.91	0.91	0.53	↑ 4.848
SK	0.94	0.92	0.94	0.75	⟷	0.94	0.97	0.53	↑ 3.628
AM	0.82	0.82	0.88	0.81	↑ 7.31	0.92	0.95	0.57	↑ 4.39
	G-Mean								
CC	0.8	0.84	0.91	0.79	↑ 14.4	0.89	0.9	0.5	↑ 13.54
GM	0.41	0.49	0.42	0.44	↑ 19.7	0.75	0.75	0.29	↑ 82.39
PH	0.85	0.88	0.87	0.88	↑ 3.07	0.9	0.9	0.25	↑5.688
SK	0.94	0.92	0.94	0.73	↑ 0.5	0.94	0.94	0.5	⟷
AM	0.85	0.88	0.87	0.87	↑ 2.35	0.94	0.94	0.29	↑ 10.58
	Balanced Accuracy								
CC	0.82	0.85	0.93	0.68	↑ 13.9	0.9	0.91	0.53	↑11.19
GM	0.58	0.6	0.58	0.48	↑ 3.63	0.76	0.76	0.54	↑ 30.55
PH	0.86	0.87	0.87	0.81	↑ 1.05	0.91	0.91	0.51	↑ 4.848
SK	0.94	0.92	0.94	0.91	⟷	0.94	0.97	0.45	↑ 3.628
AM	0.80	0.86	0.87	0.82	↑ 8.75	0.93	0.94	0.56	↑ 17.5

split between the 3rd and 4th quadrants with 3/4 and 1/1 proportionality respectively. GAN reports a maximum increase of 2%.

Therefore, following model-metric mappers are identified. VAE for Precision and F1-Score, RBM for Recall, GAN for AUC and Balanced accuracy and GAN and VAE for G-Mean. A discussion on metric-wise sample proportionality is provided in Section 8.

C. GENERATIVES + REDUCTIVES VS SYNTHETICS + REDUCTIVES

This section compares the generatives+reductives from Artifact II with the same state-of-the-art synthetic models from the previous section. Both employ IHT as majority instances reduction technique. The comparison is performed proportion wise. The objectives of this comparison being:

1) Include a majority reduction technique with both the generative and synthetic approaches. This strengthens the argument of balancing the dataset not only by augmentation but also by reduction.
2) Reinforce and improve the identified model-metric mapper. The addition of reduction technique may increase the efficiency of the model-metric and consistent findings will strengthen the argument of using deep generative models.

TABLE 4. Precision - generatives vs synthetics.

	Synthetics			Generatives			Synthetics			Generatives		
	1/4						2/4					
	GMT	KMT	SMT	VAE	GAN	RBM	GMT	KMT	SMT	VAE	GAN	RBM
CC	0.23	**0.83**	0.16	0.80	0.72	0.05	0.16	0.78	0.09	0.77	0.64	0.04
GM	0.50	**0.51**	0.41	**0.54**	0.51	0.07	0.49	0.51	0.38	0.53	0.51	0.11
PH	0.82	0.86	0.64	**0.89**	0.83	0.02	0.78	0.85	0.33	0.81	0.78	0.05
SK	0.88	**0.89**	0.88	**0.9**	0.84	0.84	0.87	0.89	0.87	0.89	0.79	0.82
AM	0.80	**0.84**	0.61	**0.88**	0.84	0.05	0.76	0.83	0.39	0.82	0.80	0.05
	3/4						4/4					
CC	0.13	0.75	0.07	**0.87**	0.62	0.04	0.12	0.76	0.06	0.85	0.58	0.02
GM	0.46	0.52	0.33	0.51	0.52	0.16	0.46	0.53	0.29	0.51	0.51	0.12
PH	0.74	**0.88**	0.28	0.82	0.81	0.4	0.73	0.83	0.23	0.86	0.75	0.18
SK	0.89	0.89	0.88	0.89	0.78	0.33	0.87	0.88	0.87	0.89	0.75	0.31
AM	0.78	0.82	0.59	0.86	0.81	0.04	0.73	0.81	0.39	0.80	0.78	0.09

TABLE 5. Recall - generatives vs synthetics.

	Synthetics			Generatives			Synthetics			Generatives		
	1/4						2/4					
	GMT	KMT	SMT	VAE	GAN	RBM	GMT	KMT	SMT	VAE	GAN	RBM
CC	0.76	0.64	0.83	0.63	0.71	**0.95**	0.8	0.67	0.85	0.66	0.79	0.92
GM	0.21	0.18	0.33	0.19	0.18	**0.97**	0.2	0.18	0.43	0.19	0.18	0.93
PH	0.74	0.74	0.84	0.74	0.76	**0.96**	0.74	0.75	0.86	0.76	0.73	0.96
SK	0.9	0.93	0.9	0.89	0.92	0.86	0.96	0.93	0.95	0.87	0.95	0.88
AM	0.72	0.74	0.81	0.73	0.75	**0.96**	0.74	0.74	0.85	0.75	0.71	0.95
	3/4						4/4					
CC	0.78	0.66	0.86	0.69	0.79	0.93	0.78	0.68	**0.88**	0.7	0.81	0.94
GM	0.22	0.17	0.49	0.18	0.18	0.96	0.24	0.17	**0.54**	0.19	0.17	0.94
PH	0.72	0.73	0.89	0.77	0.74	0.95	0.73	0.75	**0.9**	0.75	0.75	0.85
SK	0.97	0.94	0.95	0.85	0.96	**0.98**	0.96	0.92	**0.96**	0.88	0.96	0.92
AM	0.73	0.73	0.82	0.74	0.73	0.95	0.71	0.64	**0.86**	0.73	0.73	0.94

TABLE 6. F1-Score - generatives vs synthetics.

	Synthetics			Generatives			Synthetics			Generatives		
	1/4						2/4					
	GMT	KMT	SMT	VAE	GAN	RBM	GMT	KMT	SMT	VAE	GAN	RBM
CC	0.36	0.74	0.27	0.73	**0.75**	0.04	0.31	0.71	0.19	0.72	0.71	0.03
GM	0.28	0.27	0.35	0.27	0.26	0.06	0.29	0.27	0.4	0.27	0.26	0.09
PH	0.76	0.78	0.56	0.79	0.77	0.01	0.75	0.79	0.48	0.78	0.77	0.04
SK	0.9	0.91	0.89	0.89	0.89	0.42	0.92	0.92	0.92	0.89	0.87	0.42
AM	0.76	0.77	0.55	0.77	0.75	0.03	0.74	0.78	0.49	0.77	0.77	0.05
	3/4						4/4					
CC	0.22	0.71	0.13	**0.78**	0.7	0.38	0.21	0.73	0.12	0.74	0.66	0.01
GM	**0.31**	0.26	**0.41**	0.27	0.27	0.13	0.30	0.26	0.38	0.27	0.26	0.01
PH	0.72	**0.8**	0.41	**0.82**	0.76	0.28	0.73	0.79	0.38	0.79	0.75	0.14
SK	0.91	0.9	**0.92**	**0.94**	0.86	0.24	0.92	0.9	0.92	0.89	0.85	0.23
AM	0.74	**0.8**	0.42	**0.84**	0.74	0.76	0.03	0.78	0.49	0.77	0.77	0.06

3) Improve metric wise sample proportionality. As the proportion will comprise lesser majority instances, this may optimize the effective samples count.

The results are tabulated in Tables 10 to 15 and proportionality summarized in Figure 2. *Top, second best scores and proportionality quadrants use the same colours for*

TABLE 7. AUC - generatives vs synthetics.

	Synthetics			Generatives			Synthetics			Generatives		
	1/4						2/4					
	GMT	KMT	SMT	VAE	GAN	RBM	GMT	KMT	SMT	VAE	GAN	RBM
CC	0.88	0.83	0.92	0.83	0.9	0.86	0.9	0.84	0.92	0.83	0.89	0.82
GM	0.59	0.58	0.64	0.58	0.58	0.59	0.59	0.58	0.69	0.59	0.58	0.59
PH	0.87	0.87	0.91	0.87	0.88	0.5	0.87	0.88	0.92	0.87	0.88	0.6
SK	0.94	0.92	0.91	0.93	0.94	0.91	0.94	0.93	0.94	0.93	0.95	0.91
AM	0.84	0.85	0.92	0.86	0.88	0.51	0.87	0.87	0.92	0.88	0.87	0.58
	3/4						4/4					
CC	0.88	0.83	0.92	0.85	0.89	0.78	0.88	0.84	**0.93**	0.85	**0.94**	0.68
GM	0.6	0.58	0.71	0.58	0.59	0.59	0.6	0.58	**0.72**	0.58	**0.59**	0.59
PH	0.85	0.87	0.93	0.87	0.88	0.7	0.86	0.86	**0.94**	**0.88**	**0.88**	0.8
SK	0.94	0.95	0.91	0.91	0.94	0.93	0.95	0.94	**0.95**	0.92	**0.97**	0.75
AM	0.85	0.86	0.92	0.86	0.88	0.71	0.85	0.85	**0.92**	0.82	**0.93**	0.81

TABLE 8. G-Mean - generatives vs synthetics.

	Synthetics			Generatives			Synthetics			Generatives		
	1/4						2/4					
	GMT	KMT	SMT	VAE	GAN	RBM	GMT	KMT	SMT	VAE	GAN	RBM
CC	0.87	0.81	0.92	0.81	0.89	0.55	0.87	0.81	0.91	0.83	0.87	0.81
GM	0.41	0.42	0.52	0.42	0.42	0.4	0.42	0.42	0.6	0.42	0.42	0.44
PH	0.87	0.86	0.9	0.86	0.87	0.4	0.86	0.86	0.92	0.87	0.85	0.44
SK	0.95	0.95	0.95	0.93	0.95	0.90	0.93	0.95	0.96	0.93	0.95	0.91
AM	0.87	0.85	0.89	0.86	0.87	0.41	0.85	0.85	0.91	0.87	0.85	0.43
	3/4						4/4					
CC	0.88	0.83	0.91	0.83	0.88	0.76	0.88	0.83	**0.93**	0.84	**0.91**	0.79
GM	0.45	0.42	0.65	0.42	0.42	0.45	0.48	0.42	**0.69**	**0.49**	0.42	0.44
PH	0.86	0.86	0.93	0.85	0.87	0.74	0.85	0.85	**0.94**	**0.88**	0.87	0.88
SK	**0.97**	0.94	0.94	0.91	0.94	0.93	0.97	0.94	0.94	0.92	**0.97**	0.73
AM	0.86	0.85	0.93	0.85	0.86	0.75	0.84	0.84	**0.93**	**0.88**	0.87	0.87

TABLE 9. Balanced accuracy - generatives vs synthetics.

	Synthetics			Generatives			Synthetics			Generatives		
	1/4						2/4					
	GMT	KMT	SMT	VAE	GAN	RBM	GMT	KMT	SMT	VAE	GAN	RBM
CC	0.89	0.83	0.91	0.83	0.89	0.82	0.9	0.84	0.91	0.83	0.89	0.78
GM	0.58	0.58	0.63	0.58	0.58	0.59	0.58	0.58	0.67	0.59	0.58	0.59
PH	0.87	0.87	0.9	0.87	0.88	0.76	0.87	0.87	0.92	0.87	0.87	0.76
SK	0.94	0.95	0.94	0.93	0.94	0.92	0.95	0.95	0.96	0.93	0.97	0.9
AM	0.86	0.86	0.89	0.87	0.88	0.87	0.76	0.87	0.87	0.86	0.86	0.76
	3/4						4/4					
CC	0.89	0.84	0.91	0.85	0.89	0.8	0.88	0.84	**0.91**	0.85	**0.93**	0.68
GM	0.59	0.58	0.69	0.58	0.58	0.59	0.59	0.58	**0.72**	0.58	**0.61**	0.48
PH	0.86	0.86	**0.94**	**0.89**	0.87	0.88	0.86	0.87	0.94	0.87	0.87	0.81
SK	0.95	0.95	**0.96**	0.91	**0.98**	0.90	0.96	0.95	0.96	0.92	0.94	0.91
AM	0.86	0.86	**0.92**	0.88	**0.93**	0.87	0.87	0.88	0.92	0.86	0.87	0.82

analogy. Table 10 reports results on the Precision metrics. The VAE-i model leads on all 5 datasets with (0.74, 0.46,0.7,0.92,0.69). GAN-i and SMT-i trail with (0.71,0.45) and (0.67,0.9,0.67) respectively. All of the best scores are reported in the 1st quadrant where the minority-to-majority ratio is 1/7. VAE-i reports a maximum increase of 6%.

TABLE 10. Precision: Generatives + reductives vs synthetics + reductives.

	Synthetics+Reductives			Generatives+Reductives			Synthetics+Reductives			Generatives+Reductives		
	3/22 ≈ 1/7						6/19 ≈ 1/3					
	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i
CC	0.54	0.68	0.48	**0.74**	**0.71**	0.52	0.21	0.67	0.17	0.67	0.65	0.68
GM	0.44	0.43	0.43	**0.46**	**0.45**	0.07	0.37	0.38	0.34	0.38	0.37	0.11
PH	**0.67**	0.66	0.51	**0.7**	0.61	0.02	0.4	0.51	0.26	0.54	0.43	0.05
SK	**0.9**	0.89	0.89	**0.92**	0.86	0.2	0.88	0.89	0.89	0.89	0.84	0.2
AM	**0.67**	0.61	0.55	**0.69**	0.60	0.03	0.39	0.52	0.26	0.52	0.41	0.04
	9/16 ≈ 1/2						12/13 ≈ 1					
CC	0.14	0.69	0.09	0.67	0.58	0.27	0.11	0.66	0.07	0.67	0.54	0.32
GM	0.32	0.33	0.28	0.32	0.32	0.16	0.26	0.26	0.22	0.27	0.27	0.12
PH	0.42	0.5	0.21	0.57	0.41	0.03	0.38	0.51	0.18	0.54	0.43	0.18
SK	0.87	0.89	0.88	0.89	0.82	0.2	0.86	0.89	0.87	0.91	0.79	0.2
AM	0.41	0.52	0.23	0.58	0.41	0.05	0.39	0.49	0.17	0.54	0.43	0.21

TABLE 11. Recall: Generatives + reductives vs synthetics + reductives.

	Synthetics+Reductives			Generatives+Reductives			Synthetics+Reductives			Generatives+Reductives		
	3/22 ≈ 1/7						6/19 ≈ 1/3					
	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i
CC	0.81	0.8	0.84	0.8	0.83	0.51	0.79	0.8	0.86	0.8	0.81	**0.89**
GM	0.37	0.36	0.38	0.37	0.37	0.11	0.54	0.53	0.61	0.54	0.54	**0.91**
PH	0.8	0.8	0.84	0.81	0.83	0.88	0.84	0.83	0.89	0.83	0.85	**0.96**
SK	0.92	0.91	0.9	0.9	0.93	0.98	0.95	0.93	0.96	0.92	0.94	**0.99**
AM	0.79	0.79	0.83	0.82	0.84	0.88	0.88	0.84	0.88	0.85	0.86	**0.92**
	9/16 ≈ 1/2						12/13 ≈ 1					
CC	0.8	0.81	0.86	0.8	0.83	0.68	0.8	0.81	**0.87**	0.8	0.82	0.65
GM	0.61	0.6	0.68	0.61	0.6	0.91	0.64	0.64	**0.72**	0.64	0.64	0.91
PH	0.83	0.83	0.92	0.82	0.84	0.90	0.85	0.83	**0.93**	0.82	0.82	0.91
SK	0.97	0.93	0.96	0.91	0.94	0.95	0.97	0.91	**0.98**	0.95	0.97	0.97
AM	0.82	0.82	0.84	0.81	0.83	0.88	0.86	0.84	**0.89**	0.81	0.81	0.89

TABLE 12. F1-Score: Generatives + reductives vs synthetics + reductives.

	Synthetics+Reductives			Generatives+Reductives			Synthetics+Reductives			Generatives+Reductives		
	3/22 ≈ 1/7						6/19 ≈ 1/3					
	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i
CC	0.64	0.73	0.61	0.72	0.75	0.25	0.34	**0.74**	0.28	0.73	**0.76**	0.38
GM	0.41	0.4	0.4	0.41	0.41	0.04	0.42	0.42	0.42	**0.44**	**0.43**	0.09
PH	0.72	**0.73**	0.63	**0.75**	0.71	0.02	0.56	0.64	0.42	0.66	0.58	0.04
SK	0.91	0.91	0.92	0.9	0.9	0.16	0.91	**0.92**	0.92	**0.93**	0.87	0.16
AM	0.56	0.62	0.41	0.67	0.58	0.04	0.65	0.70	0.61	**0.76**	**0.74**	0.05
	9/16 ≈ 1/2						12/13 ≈ 1					
CC	0.22	0.73	0.15	0.73	0.65	0.19	0.19	0.73	0.13	0.73	0.73	0.21
GM	0.41	0.41	0.39	0.41	0.41	0.13	0.39	0.39	0.35	0.39	0.39	0.11
PH	0.55	0.64	0.35	0.66	0.57	0.02	0.51	0.65	0.31	0.67	0.58	0.15
SK	0.91	0.91	0.91	0.9	0.86	0.17	0.91	0.9	0.91	0.9	0.87	0.17
AM	0.54	0.64	0.34	0.67	0.58	0.05	0.49	0.65	0.32	0.68	0.55	0.21

RBM-i model clearly leads on all datasets using Recall metrics with (0.89,0.91,0.96,0.99,0.92) as shown in Table 11. SMT-i trails with (0.87,0.72,0.93,0.98,0.89). The top results are reported in 2nd quadrant where the minority-to-majority ratio is 1/3. RBM-i reports a maximum increase of 19%. Table 12 reports results on F1-Score. Similar to the Precision

TABLE 13. AUC: Generatives + reductives vs synthetics + reductives.

	Synthetics+Reductives			Generatives+Reductives			Synthetics+Reductives			Generatives+Reductives		
	$3/22 \approx 1/7$						$6/19 \approx 1/3$					
	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i
CC	0.9	0.9	0.92	0.9	0.91	0.5	0.89	0.9	0.91	0.9	0.9	0.58
GM	0.7	0.69	0.71	0.7	0.7	0.52	0.73	0.73	0.75	0.73	0.73	0.5
PH	0.9	0.9	0.93	0.9	0.91	0.54	0.91	0.91	0.93	0.91	0.91	0.5
SK	0.95	0.95	0.96	0.94	0.95	0.55	0.95	0.95	0.96	0.94	0.94	0.56
AM	0.89	0.88	0.91	0.91	0.90	0.56	0.90	0.90	0.91	0.91	0.92	0.55
	$9/16 \approx 1/2$						$12/13 \approx 1$					
CC	0.89	0.9	0.92	0.9	0.91	0.6	0.9	0.91	**0.92**	0.9	**0.93**	0.55
GM	0.75	0.74	0.76	0.75	0.75	0.5	0.76	0.75	**0.77**	0.76	**0.79**	0.5
PH	0.91	0.91	0.94	0.91	0.92	0.5	0.91	0.91	**0.95**	0.91	**0.92**	0.53
SK	0.95	0.95	0.96	0.94	0.94	0.51	0.95	0.94	**0.96**	0.94	**0.97**	0.53
AM	0.91	0.91	0.92	0.91	0.91	0.51	0.91	0.91	**0.93**	0.92	**0.95**	0.57

metrics, the generative+reductive models leads the synthetic counterparts on all datasets. VAE-i reports the best score on 4 datasets with (0.44, 0.75, 0.93,0.76) and GAN-i on 1 with (0.76). KMT-i trails with (0.74,0.42,0.73,0.92,0.70). The best scores are reported in 2nd quadrant where the minority-to-majority ratio is 1/3. VAE-i reports a maximum increase of 2%. Table 13 reports the results on AUC metrics. GAN-i clearly leads on 4 datasets and is comparable on 1 with (0.93,0.79,0.92,0.97,0.95), SMT-i trails with (0.92,0.77,0.95,0.96,0.93). The best scores are reported in the 4th quadrant where the minority nearly equals majority. A maximum increase of 2% is reported by GAN-i. Table 14 show results on G-Mean metric where VAE-i and GAN-i closely follow. The scores are reported in 4th quadrant. The results on Balanced accuracy metrics are reported in Table 15. GAN-i lead on 3 and SMT-i on 2 datasets with scores (0.92,0.77,0.93,0.97,0.94) and (0.93,0.76,0.94,0.96,0.92) respectively. The majority of the best scores are reported in 4th quadrant where minority nearly equals majority. GAN-i reports a maximum increase of 2%.

However, the method in this section generates and reduces the minority and majority respectively, the findings are consistent with the previous section where the minority is generated only. To elaborate, similar model-metric mapping is identified. VAE-i for Precision and F1-Score, RBM-i for Recall, GAN-i for AUC and Balanced accuracy and VAE-i and GAN-i for G-Mean. As for the metric-wise sample proportionality a discussion is provided in Section 8.

D. COMPUTATIONAL EFFICIENCY

As for training efficiency, the generative models scale linearly as opposed to the synthetic counterparts which scale in orders of multiple. To elaborate, if there are m datasets and each is to be augmented using n proportions, than the generative models merely require m while the synthetic require $m \times n$ training time. As this work uses 5 datasets each with 4 proportions, the training time for generative models is 5 while for synthetic models is 5×4. This linear order training efficiency makes the generative models a stronger candidate.

VIII. MMM IN MOTION

This section sets the proposed methodology in motion. MMM launches a six pronged attack to neutralize class imbalance from six frontiers as shown in figure 3. The objective is to come up with a data driven and industry neutral class imbalance solution. The motion is set as follows:

- The optimum minority-to-majority ratio against specific metric is identified. The sensitivity of the metric to varying degree of proportionality is elaborated. The model-metric mapping is established. The rationalization is strengthened by observing these against both categories.
- An inter-comparison of the leading models from each category is performed establishing a rank-order preference. MMM, finally recommends this rank-order based model-metric mapping along with the optimum minority-to-majority proportionality.

1) *Precision metrics* require a low minority-to-majority ratio. The reason being prime results are reported in 1st quadrant with sample proportionality 1/4 and 1/7 respectively, Figure 2. The scores drop in higher quadrants where minority representation increases, endorsing the sensitivity of the metric to proportionality, refer Tables 4, 10. VAE and VAE-i lead over competing models in respective categories with mentioned proportionality. This highlights high representational strength of VAE generated instances against precision. The reduction in false positives leading to high Precision can be attributed to these instances, marking the suitability of the VAE model against the metric.

Recommendation - VAE from Artifact I, 1/4 proportionality An inter-category comparison between the two leaders observes VAE surpasses VAE-i showing the

TABLE 14. G-Mean: Generatives + reductives vs synthetics + reductives.

	Synthetics+Reductives			Generatives+Reductives			Synthetics+Reductives			Generatives+Reductives		
	3/22 ≈ 1/7						6/19 ≈ 1/3					
	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i
CC	0.9	0.89	0.91	0.89	0.91	0.75	0.88	0.89	0.92	0.89	0.9	0.53
GM	0.6	0.59	0.6	0.6	0.6	0.3	0.69	0.68	0.72	0.69	0.69	0.25
PH	0.9	0.9	0.93	0.89	0.91	0.48	0.91	0.91	0.93	0.91	0.92	0.1
SK	0.96	0.95	0.96	0.94	0.96	0.5	0.95	0.95	0.96	0.94	0.94	0.5
AM	0.90	0.90	0.92	0.88	0.91	0.51	0.89	0.91	0.90	0.91	0.91	0.21
	9/16 ≈ 1/2						12/13 ≈ 1					
CC	0.89	0.9	0.92	0.89	0.91	0.53	0.89	0.9	**0.92**	0.89	**0.92**	0.5
GM	0.72	0.71	0.74	0.72	0.72	0.21	0.72	0.74	**0.76**	**0.75**	**0.75**	0.29
PH	0.91	0.91	**0.95**	0.9	0.91	0.1	0.92	0.91	0.92	**0.93**	**0.93**	0.25
SK	**0.96**	0.95	0.96	0.94	0.94	0.5	0.96	0.94	**0.97**	0.94	0.94	0.5
AM	0.91	0.91	0.93	0.90	0.89	0.21	0.91	0.91	0.92	**0.94**	**0.94**	0.29

TABLE 15. Balanced accuracy - generatives + reductives vs synthetics + reductives.

	Synthetics+Reductives			Generatives+Reductives			Synthetics+Reductives			Generatives+Reductives		
	3/22 ≈ 1/7						6/19 ≈ 1/3					
	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i	GMT-i	KMT-i	SMT-i	VAE-i	GAN-i	RBM-i
CC	0.9	0.9	0.92	0.9	0.91	0.78	0.89	0.9	0.92	0.9	0.9	0.6
GM	0.67	0.67	0.67	0.67	0.67	0.51	0.72	0.72	0.74	0.72	0.72	0.5
PH	0.9	0.9	0.93	0.9	0.92	0.56	0.91	0.91	0.93	0.91	0.92	0.5
SK	**0.96**	0.95	**0.96**	0.94	0.95	0.4	0.96	0.95	0.96	0.94	0.94	0.1
AM	0.89	0.89	0.91	0.90	0.91	0.55	0.90	0.90	0.91	0.92	0.92	0.54
	9/16 ≈ 1/2						12/13 ≈ 1					
CC	0.89	0.9	0.92	0.9	0.91	0.55	0.9	0.91	**0.93**	0.9	**0.92**	0.53
GM	0.75	0.75	0.75	0.75	0.75	0.55	0.75	0.75	**0.75**	**0.75**	**0.77**	0.54
PH	0.91	0.91	0.93	0.91	0.92	0.53	0.92	0.91	**0.94**	0.91	**0.93**	0.51
SK	0.96	0.95	0.96	0.94	0.94	0.2	0.96	0.94	0.96	0.94	**0.97**	0.45
AM	0.89	0.89	0.91	0.92	0.92	0.52	0.90	0.91	0.92	**0.93**	**0.94**	0.56

generated instances are even more expressive independently, Tables 4, 10. Thus, for high precision, MMM recommends VAE from Artifact I followed by VAE-i from Artifact - II with the mentioned proportionality.

2) *Recall metrics* similarly require a low minority-to-majority ratio as prime results are reported in the 1st and 2nd quadrant with sample proportionality 1/4 and 1/3 respectively, Figure 2. Scores decline in higher quadrants as minority representation is increased, cementing the sensitivity of the metric to proportionality, Tables 5, 11. RBM and RBM-i are top performers in their respective categories with the mentioned proportionality. High representational strength of RBM generated instances against can be confirmed as synthetic SMT trails in both categories and that also using high proportion of minority instances. The reduction in false negatives leading to high Recall can be attributed to these instances, marking the suitability of the RBM model against the metric.

Recommendation - RBM from Artifact I, 1/4 proportionality An inter-category comparison between two foremost show that RBM further leads over RBM-i, Tables 5, 11. This shows that generative instances from RBM are independently more expressive than being combined with reductive technique as the later not only require more instances but also an equivalent reduction of the majority class. Thus, for high recall, MMM recommends RBM from Artifact I followed by RBM-i from Artifact - II with the mentioned sample proportionality.

3) *F1-Score* require a moderate minority-to-majority ratio as highest scores are reported in the 3rd and 2nd quadrant with 3/4 and 1/3 sample proportionality respectively, Figure 2. Sensitivity of the metric to proportionality can be observed as scores drop when minority-to-majority ratio is shifted to either extreme, Tables 6, 12. VAE and VAE-i exceed in their respective categories. A modest representation strength of VAE generated instances is observed against the F1-Score. The homogeneity between false positives and negatives

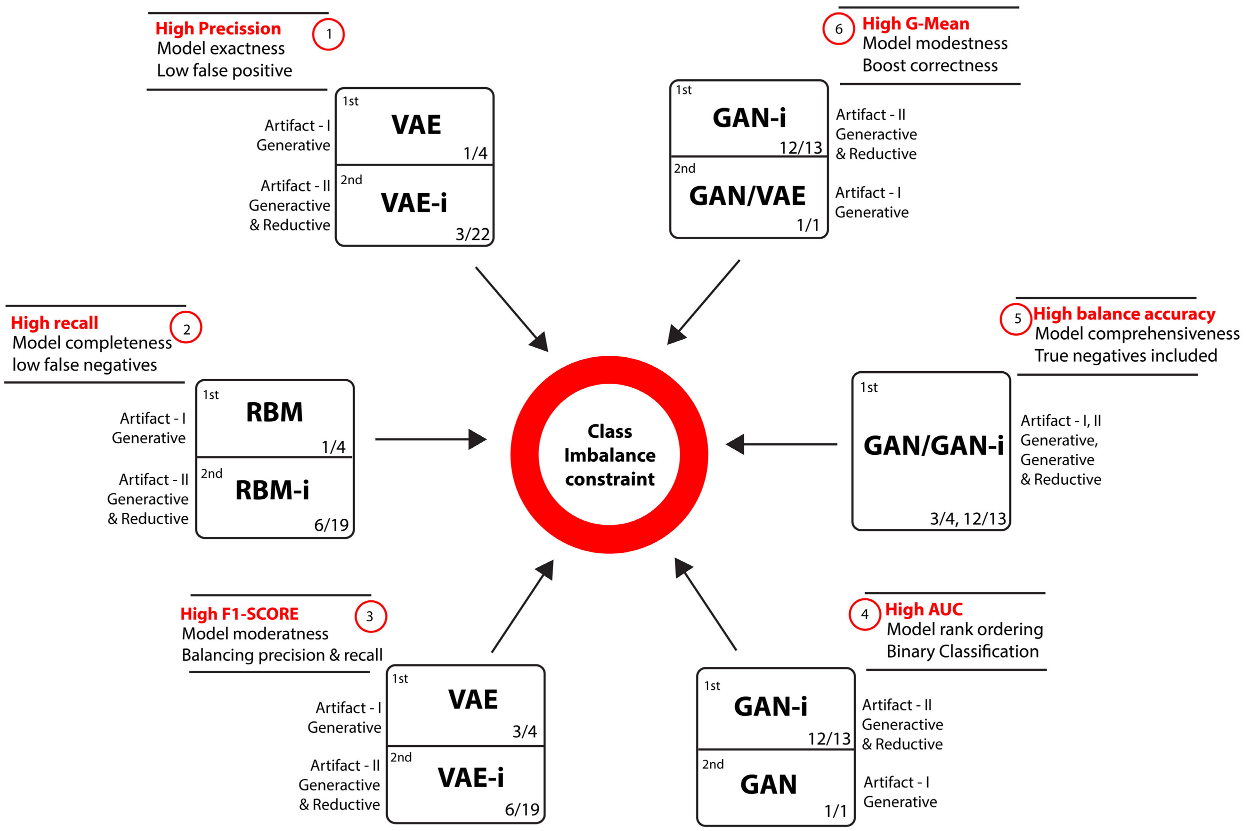


FIGURE 3. MMM in motion - A six pronged attack on imbalance.

leading to high F1-Score can be attributed to these instances, marking the suitability of the VAE model against the metric.

Recommendation - VAE from Artifact I, 3/4 proportionality Comparing the two leaders from each category show VAE surpasses VAE-i,Tables 6, 12. This confirms a moderate representation of minority class is preferred over a restricted one. Therefore for high F1-Score, MMM recommends VAE as 1st and VAE-i as 2nd model with proportionality of 3/4 and 1/3 respectively.

4) *AUC metrics* require a high or near equal minority-to-majority ratio as the top scores being reported in 4th quadrant with sample proportionality 12/13 and 1/1 respectively, Figure 2. The metric is highly sensitive to proportionality as low scores are observed until near equilibrium between the two classes is achieved, Tables 7, 13. GAN-i and GAN lead on multiple datasets in their respective categories. This shows high expressiveness of GAN generated instances over other models against AUC. The enhanced grading predictions capacity leading to high AUC can be attributed to these instances, marking the suitability of the GAN model against the metric.

Recommendation - GAN-i from Artifact II, 12/13 proportionality: An inter-category comparison of the two prime performers observe GAN-i excels over GAN, Tables 7, 13. This shows that for AUC metric, GAN instances are more expressive when a near equilibrium of both classes is maintained but with high count. Therefore for high AUC, MMM endorse GAN-i as 1st and GAN as 2nd model with proportionality 12/13 and 1/1 respectively.

5) *G-Mean metrics* Similar to AUC, G-Mean require a high or near equal minority-to-majority ratio as the top scores being reported in 4th quadrant with sample proportionality 1/1 and 12/13 respectively, Figure 2. Sensitivity to proportionality is evident as low scores are reported with low minority counts, Tables, 8, 14. GAN, VAE and GAN-i deliver near comparable performance against foremost models. To enjoy expressiveness, both synthetic and generated instances require a near equal presence of the opposite class with high count. Balancing the dataset is attributed to these instances which increase modestness and leads to high G-Mean, marking the suitability of GAN and VAE models against the metric.

Recommendation - G-Mean, GAN-i from Artifact II, 12/13 proportionality: An inter-category comparison between deep models establishes lead of GAN-i over GAN and VAE, Tables 8, 14. MMM, for high

G-Mean, recommends GAN-i together with SMT as the 1st model followed by GAN and VAE with proportionality 12/13 and 1/1 respectively.

6) *Balanced accuracy metric* requires a moderate to high ratio. The reason being metric reports highest score in the 4th and a split between 3rd and 4th quadrants with proportionality 12/13, 3/4 and 1/1 respectively, Figure 2. The metric is observed to have medium sensitivity as low scores fall in lower quadrants, Tables 9, 15. GAN and GAN-i lead on multiple datasets in their categories. GAN generated instances enjoy high expressive strength against balanced accuracy. The equilibrium attained between true positives and negatives is attributed to these instances which increase comprehensiveness and leads to high Balanced accuracy, marking the suitability of the GAN model against the metric.

   Recommendation - Balanced accuracy, GAN,GAN-i from Artifact I,II, 3/4,1/1,12/13 proportionality: Intercomparison between the foremost models show near equivalent performance, Tables 9, 15. Hence, MMM recommends both GAN and GAN-i using mentioned proportionality.

IX. CONCLUSION

The proposed MMM methodology, covers research gap in class imbalance domain by building on two concepts. First, the authors are of the view that metrics being distinct in their formulation and usage are also sensitive to data proportions but with varying degree. An effective proportionality for one metric may be not be suitable for the other. Therefore metric wise proportionality calibration is required. Second, a highly suitable model on one metric may be less suitable on the other. So, an informed model selection is required. Though, deep models are known to have strong generative capabilities, but their inherent architectural and algorithmic variation also makes a strong case for precise candidate selection. MMM, formulated on these concepts, conclude the following:

1) Optimal model-metric mapping identified and 1st, 2nd recommendation proposed. These are, Precision and F1-Score: VAE, VAE-i, Recall: RBM, RBM-i, AUC and G-Mean: GAN-i, GAN/VAE and Balanced accuracy: GAN/GAN-i.
2) Metric wise optimum minority-to-majority proportionality is calibrated on both Augmentation and Augmentation + Reduction categories. These are, Precision: 1/4, 1/7, Recall: 1/4,1/3, F1-Score 3/4,1/3, AUC and G-Mean: 1/1,12/13, and Balanced accuracy: 3/4, 12/13.
3) The proposed deep models, outperform synthetic counterparts on Precision, Recall, and F1-Score, AUC and Balanced accuracy on both categories. The maximum % increase are Precision: 4,6, Recall: 43, 19, F1-Score: 5, 2, AUC: 2, 2 and Balanced accuracy: 2, 2.
4) The proposed deep models comprehensively surpass baselines on all 6 metrics on both categories. The maximum % increase are Precision: 1.71,1.16, Recall: 452.9,435.29, F1-Score: 7.46,51.67, AUC: 15.2,30.55, G-Mean:19.7,82.39, Balanced accuracy: 13.9,30.55.
5) The proposed deep models are computational efficient as these scale linearly in m as opposed to the synthetic counterparts which scale in orders of multiple in $m \times n$.

REFERENCES

[1] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[2] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Inf. Sci.*, vol. 512, pp. 1214–1233, Feb. 2020.

[3] R. Janbandhu, S. Begum, and N. Ramasubramanian, "Credit card fraud detection," in *Computing in Engineering and Technology*. Singapore: Springer, 2020, pp. 225–238.

[4] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," 2020, *arXiv:2008.09202*. [Online]. Available: http://arxiv.org/abs/2008.09202

[5] M. Koziarski, B. Krawczyk, and M. Woźniak, "Radial-based oversampling for noisy imbalanced data classification," *Neurocomputing*, vol. 343, pp. 19–33, May 2019.

[6] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci.*, vol. 465, pp. 1–20, Oct. 2018.

[7] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci.*, vol. 501, pp. 118–135, Oct. 2019.

[8] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Inf. Sci.*, vol. 479, pp. 448–455, Apr. 2019.

[9] Y.-J. Zheng, X.-H. Zhou, W.-G. Sheng, Y. Xue, and S.-Y. Chen, "Generative adversarial network based telecom fraud detection at the receiving bank," *Neural Netw.*, vol. 102, pp. 78–86, Jun. 2018.

[10] Y. Guo, W. Liao, Q. Wang, L. Yu, T. Ji, and P. Li, "Multidimensional time series anomaly detection: A GRU-based Gaussian mixture variational autoencoder approach," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 97–112.

[11] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," 2019, *arXiv:1904.06991*. [Online]. Available: http://arxiv.org/abs/1904.06991

[12] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR*, 2019, pp. 1–35.

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4401–4410.

[14] H. Tingfei, C. Guangquan, and H. Kuihua, "Using variational auto encoding in credit card fraud detection," *IEEE Access*, vol. 8, pp. 149841–149853, 2020.

[15] Z. Islam, M. Abdel-Aty, Q. Cai, and J. Yuan, "Crash data augmentation using variational autoencoder," *Accident Anal. Prevention*, vol. 151, Mar. 2021, Art. no. 105950.

[16] W. Dai, K. Ng, K. Severson, W. Huang, F. Anderson, and C. Stultz, "Generative oversampling with a contrastive variational autoencoder," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 101–109.

[17] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," 2018, *arXiv:1806.00035*. [Online]. Available: http://arxiv.org/abs/1806.00035

[18] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," 2018, *arXiv:1806.03384*. [Online]. Available: http://arxiv.org/abs/1806.03384

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*. [Online]. Available: http://arxiv.org/abs/1706.08500

[20] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[21] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Inf. Sci.*, vol. 409, pp. 17–26, Oct. 2017.

[22] A. Oussidi and A. Elhassouny, "Deep generative models: Survey," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–8.

[23] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving imbalanced dataset classification using oversampling and gradient boosting," in *Proc. 5th Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2019, pp. 217–222.

[24] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong, "Classification of imbalanced data by oversampling in kernel space of support vector machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4065–4076, Sep. 2018.

[25] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 935–942.

[26] A. D'Addabbo and R. Maglietta, "Parallel selective sampling method for imbalanced and large data classification," *Pattern Recognit. Lett.*, vol. 62, pp. 61–67, Sep. 2015.

[27] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 159–166.

[28] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, pp. 1653–1672, May 2015.

[29] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[30] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data-recommendations for the use of performance metrics," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 245–251.

[31] P. A. Flach, J. Hernández-Orallo, and C. F. Ramirez, "A coherent interpretation of AUC as a measure of aggregated classification performance," in *Proc. ICML*, 2011, pp. 1–8.

[32] V. García, R. A. Mollineda, and J. S. Sánchez, "On the k-NN performance in a challenging scenario of imbalance and overlapping," *Pattern Anal. Appl.*, vol. 11, nos. 3–4, pp. 269–280, Sep. 2008.

[33] D. A. Cieslak and N. V. Chawla, "Start globally, optimize locally, predict globally: Improving performance on imbalanced data," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 143–152.

[34] N. Japkowicz and R. Holte, "Workshop report: AAAI-2000 workshop on learning from imbalanced data-sets," *AI Mag.*, vol. 22, no. 1, pp. 127–136, 2000.

[35] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Syst. Appl.*, vol. 91, pp. 464–471, Jan. 2018.

[36] Y. Zhang, "Deep generative model for multi-class imbalanced learning," M.S. thesis, NOVA Inf. Manage. School, Universidade Nova de Lisboa, Lisbon, Portugal, 2018. [Online]. Available: https://digitalcommons.uri.edu/theses/1277

[37] M. Zięba, J. M. Tomczak, and A. Gonczarek, "RBM-SMOTE: Restricted Boltzmann machines for synthetic minority oversampling technique," in *Proc. 7th Asian Conf. ACIIDS*, Bali, Indonesia, Mar. 2015, pp. 377–386.

[38] I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst. Appl.*, vol. 46, pp. 405–416, Mar. 2016.

[39] G. Douzas and F. Bacao, "Self-organizing map oversampling (SOMO) for imbalanced data set learning," *Expert Syst. Appl.*, vol. 82, pp. 40–52, Oct. 2017.

[40] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DBSMOTE: Density-based synthetic minority over-sampling technique," *Appl. Intell.*, vol. 36, no. 3, pp. 664–684, Apr. 2012.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[42] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*. El Segundo, CA, USA: Association for the Advancement of Artificial Intelligence, 2005, pp. 878–887.

[43] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, Apr. 2011.

[44] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. 7th Eur. Conf. Princ. Pract. Knowl. Discovery Databases*, Dubrovnik, Croatia, 2003, pp. 107–119.

[45] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[46] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: A case study involving information extraction," in *Proc. Workshop Learn. Imbalanced Datasets*, 2003, pp. 1–7.

[47] M. Kubat and S. Matwin, "Addressing the course of imbalanced training sets: One-sided selection," in *Proc. ICML*, 1997, pp. 179–186.

[48] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 97, pp. 769–772, Jul. 1997.

[49] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Mach. Learn.*, vol. 95, no. 2, pp. 225–256, May 2014.

[50] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*. New York, NY, USA: Springer, 2009, pp. 875–886.

[51] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, 2004.

[52] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[53] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[54] G. E. A. P. A. Batista, B. Bazzan, and M. Monard, "Balancing training data for automated annotation of keywords: A case study," in *Proc. WOB*, 2003, pp. 10–18.

[55] *The National Security Agency: Missions, Authorities, Oversight and Partnerships*, Nat. Secur. Agency, Fort Meade, MD, USA, 2003.

[56] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[57] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*. [Online]. Available: http://arxiv.org/abs/1703.10717

[58] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[59] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[60] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[61] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2794–2802.

[62] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 1133–1141.

[63] A. Roberts, J. Engel, and D. Eck, "Hierarchical variational autoencoders for music," in *Proc. NIPS Workshop Mach. Learn. Creativity Design*, 2017, pp. 1133–1141.

[64] C. K. Fisher, A. M. Smith, and J. R. Walsh, "Boltzmann encoded adversarial machines," 2018, *arXiv:1804.08682*. [Online]. Available: http://arxiv.org/abs/1804.08682

[65] K. H. Cho, T. Raiko, and A. Ilin, "Gaussian–Bernoulli deep Boltzmann machine," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–7.

[66] M. R. Smith, "An empirical study of instance hardness," Ph.D. dissertation, Dept. Inf. Comput. Sci., School Sci. Aalto Univ., Espoo, Finland, 2010.

[67] A. H. Peterson and T. R. Martinez, "Estimating the potential for combining learning models," in *Proc. ICML Workshop Meta-Learn.*, 2005, pp. 68–75.

[68] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[69] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[70] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[72] L. K. Saul and T. S. Roweis. (2000). *An Introduction to Locally Linear Embedding*. [Online]. Available: http://www.cs.toronto.edu/~roweis/lle/publications.html

[73] *Imagenet Classification With Deep Convolutional Neural Networks*. Accessed: 2012. [Online]. Available: https://www.kaggle.com/c/GiveMeSomeCredit

[74] R. Caruana, T. Joachims, and L. Backstrom, ''KDD-cup 2004: Results and analysis,'' *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 2, pp. 95–108, Dec. 2004.

[75] R. Rossi and N. Ahmed, ''The network data repository with interactive graph analytics and visualization,'' in *Proc. AAAI Conf. Artif. Intell.*, 2015, vol. 29, no. 1, pp. 1–2.

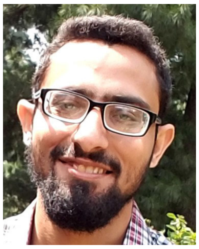
[76] *KDD-Cup 2004: Results and Analysis*. Accessed: Dec. 1, 2004. [Online]. Available: https://github.com/aaxwaz/Fraud-detection-using-deep-learning/tree/master/auto-encoder

[77] *The Network Data Repository With Interactive Graph Analytics and Visualization*. Accessed: Mar. 4, 2015. [Online]. Available: https://github.com/aaxwaz/Fraud-detection-using-deep-learning/tree/master/rbm

[78] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, ''Modeling tabular data using conditional GAN,'' 2019, *arXiv:1907.00503*. [Online]. Available: http://arxiv.org/abs/1907.00503

[79] G. Douzas, F. Bacao, and F. Last, ''Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,'' *Inf. Sci.*, vol. 465, pp. 1–20, Oct. 2018.

BEHROZ MIRZA received the M.S. (C.S.) degree from SZABIST, Karachi, Pakistan, in 2014. He is currently pursuing the Ph.D. degree in computer science with Dr. Tahir Q. Syed with the National University of Computer and Emerging Sciences, Karachi. His research interests include machine learning, deep learning, theoretical computer science, and turning machines.



DANISH HAROON is currently pursuing the M.S. degree in data science with the National University of Computer and Emerging Sciences, Karachi. His research interests include data science and machine learning.

BEHRAJ KHAN received the M.S. (C.S.) degree from the National University of Computer and Emerging Sciences, Islamabad, Pakistan, in 2016. He is currently pursuing the Ph.D. degree in computer science with Dr. Tahir Q. Syed with the National University of Computer and Emerging Sciences, Karachi, Pakistan. His research interests include machine learning, deep learning, and natural language processing.

ALI PADHANI received the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Karachi, in 2018. His research interests include data science and machine learning.

TAHIR Q. SYED received the Ph.D. degree in computer vision from Université Paris-Saclay, France. He currently lectures at the Institute of Business Administration, Pakistan. His research interests include machine learning, computer vision, and statistical inference.