Synthesizing Class Labels for Balanced and Highly Imbalanced Cognition Data

Robert K.L. Kennedy and Taghi M. Khoshgoftaar
Florida Atlantic University
Boca Raton, Florida 33431
rkennedy@fau.edu, khoshgof@fau.edu

***Abstract*—In machine learning, assembling labeled training datasets often incurs significant costs and requires expert human annotation, which is essential for effectively training supervised models. A vast majority of newly generated data is unlabeled, a situation that poses significant challenges in fields such as medical diagnosis where precise and timely labeling is critical for the early detection of cognitive conditions. In this work, we employ an unsupervised method in a novel context to synthesize class labels for two cognitive datasets derived from publicly available survey data in the Health and Retirement Study (HRS). To assess the quality and usability of the newly generated labels, we train six supervised classifiers and evaluate their classification performance. Our results indicate that the synthesized labels are of high enough quality and enable effective classifier training. Furthermore, these classifiers outperform a baseline learner on both balanced and highly imbalanced cognition datasets, as evidenced by their area under the precision-recall curve (AUPRC) scores. The results also demonstrate that the synthesized labels yield higher AUPRC scores on balanced data.**

***Index Terms*—label synthesis, unsupervised learning, class imbalance, cognition, machine learning**

I. Introduction

A growing elderly population is presenting nations worldwide with significant challenges stemming from age-related cognitive impairments, notably dementia [1]. This trend adversely affects the aging population as well as global economies [2]. A 2010 study estimated $604 billion was spent on dementia in just the United States [2]. Dementia is not a specific disease but a collective term for conditions that are characterized by a decline in cognitive function that interferes with daily life. It is a significant worldwide problem, with its prevalence increasing, and it is ranked as the seventh leading cause of death globally [3]. An estimated 153 million people will show signs of dementia by 2050, globally [4]. Machine learning has shown promise in early detection of dementia using various different types of datasets such as neuroimaging, speech analysis, and survey-based datasets like the Health and Retirement Study (HRS) datasets [5], [6]. The HRS is conducted by the University of Michigan and sponsored by the National Institute on Aging (grant number NIA U01AG009740).

The HRS is a longitudinal study that collects valuable data from participants, for various aspects of cognitive decline and other aspects of aging [7]. Some of the HRS data is publicly available and other portions of the data are restricted. Though our work focuses on HRS data that is publicly available, the data still provides valuable insights into age-related cognitive decline and early signs of dementia. While supervised machine learning methods are the focus of much of the current research in the area, the importance of unsupervised approaches remains significant. By default, most newly created data are unlabeled [8]. This presents a unique challenge to machine learning practitioners since machine learning models trained with large amounts of high-quality labeled data yield good results; however, creating this high-quality labeled data is both difficult and costly [9]. Unsupervised learning methods can utilize unlabeled data, offering the key benefit of eliminating the need for manual class labeling, which is often susceptible to errors [9].

Another challenge for machine learning is class imbalance, which is data that has a significant difference in class representation. This can be found across multiple domains such as fraud detection [10], [11], network intrusion data [12], healthcare data [13], and cognition data [14]. Class imbalance can cause a machine learning model to become biased toward the class with more instances [15] which results in negatively impacted performance. There are two main approaches for dealing with class imbalance: data-level and algorithm-level. The former are methods that modify the dataset before training in some manner, such as with random under sampling (RUS) or synthetic minority over-sampling (SMOTE). The latter are methods that rely on improving the machine learning model itself to be more robust to class imbalance. The unsupervised labeling method we employ in our work addresses the challenges of class imbalance in a hybrid fashion and has been shown to outperform Isolation Forest (IF), a popular unsupervised anomaly detection method, in highly imbalanced data [10].

In this work, we apply our label synthesizing method to both balanced and imbalanced cognition datasets. Our method employs an unsupervised neural network to produce an error metric for each instance. To the best of our knowledge, this is the first work to examine the effects of an unsupervised class labeling technique on balanced and imbalanced cognition data. From the errors calculated by the approach, we create new labels. To effectively measure the quality of the newly created class labels, we train six supervised classifiers on the new labels and measure their classification performance. Importantly, we train the classifiers on the synthesized class labels and test their performance on the original class labels

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00017

that are part of the original datasets, i.e., the original class labels are only used to measure performance and not used in model training or label synthesis. In an unsupervised scenario, either an unsupervised model, such as IF, or a supervised classifier trained on synthetic labels can be used. Our results show that the supervised classifiers trained on synthetic labels outperform the baseline unsupervised method. Comparing supervised classification performance using synthesized labels versus original labels is out of scope of this paper and would be possible future work.

II. Related Work

Alternative methods for detecting dementia, or other cognitive decline issues, involve using data sources beyond survey-based longitudinal studies, such as neuroimaging from magnetic resonance imaging (MRI) scans. Machine learning models, including support vector machines (SVM) [16], have proven to be effective in detecting early signs of cognitive decline through analysis of MRI scan data. Machine learning models have been shown to be more effective than manual assessments by medical specialists [16]. Nonetheless, MRI imaging is not readily available worldwide and comes with significant cost hurdles [17]. For that reason, researching predictive machine learning techniques aimed at alternative data sources proves beneficial.

Other machine learning techniques focus on detecting cognitive decline using other types of data sources, such as speech analysis datasets. Studies have shown that vocabulary can be used as an indicator that shows a deterioration in semantic memory, language, and the perceptual process [18], [19]. This type of approach is similar to an MRI in that it is non-invasive but has the advantage of not needing specialized personnel, laboratory equipment, and is less cost prohibitive. Neural networks have been used to effectively analyze the audio data to detect Alzheimer's disease [20], as well as bag of words [19]. However, these approaches require the use of labeled datasets, significantly differing from our approach.

Approaches that use longitudinal data, which is data that consists of repeated observations of the same subject over time [21], are able to capture complex patterns and temporal dynamics in the cognition and healthcare domains. Seligman et al. [22] explore various machine learning approaches, including linear regression, random forests, and small feed forward neural networks, for studying social factors of health using longitudinal type data, namely HRS. They show that the more complicated neural networks perform best when analyzing chronic disease HRS data. Though their work is similar to ours in that we both use HRS data, their work does not look at cognition data, their supervised approaches use the labels provided by the HRS data source, and they do not consider highly imbalanced data. Other works used unsupervised methods, such as principal component analysis (PCA) to analyze HRS data. Langavant et al., in [23], use PCA to cluster partially labeled data and compare members of the clusters with and additional source of dementia probabilities for the same instances to make a classification, i.e., their work still requires some of the data to be labeled, whereas our work does not require any class labels.

Several studies have used data from the HRS beyond survey data, including genetic information, blood test results, and other non-publicly available data [23], [24]. Our work differs from this in that we use publicly available HRS survey data which can be collected remotely, without needing medical specialists to run specialized tests. Additionally, our work differs from the existing literature in that our approach is driven by unsupervised labeling, and we evaluate our work on both balanced and highly imbalanced data.

Following our literature review, it became apparent that our automated class labeling technique for highly imbalanced datasets is a pioneering effort in the field. Due to the novelty of the research, directly comparable studies and methodologies are scarce. Nevertheless, we believe that discussing these indirectly related works adds a relevant context and highlights the uniqueness of our work.

III. Methodology

Our methodology introduces a novel way to synthesize binary class labels, originally developed for datasets with significant class imbalances [10]. In this paper, we test and evaluate its effectiveness on data that is class balanced. The labeling approach uses an autoencoder [25] to effectively learn from the features of the datasets and once trained, the autoencoder calculates an error metric for each instance. The instances are sorted from highest error to lowest error. The premise for sorting is instances that produce a higher reconstruction error from the autoencoder are more likely to be in the minority class. Instances with relatively lower errors are more likely to be in the majority class. After sorting instances based on their reconstruction error, they are now in a state such that instances toward the top are more likely to be in one class, and the instances nearer the bottom are more likely to be in the other.

An error threshold must be chosen so that instances with higher errors are labeled as positive and negative otherwise. In this work, positive instances are instances in the minority class, and negative instances are instances in the majority class. Initially, we establish an error threshold using domain expertise and insights into the characteristics of the data population. Instances that are nearest to this threshold are not strongly indicated as either one class or the other. A portion of the instances just below the threshold are labeled as positive. This is done to add diversity to the minority instance group which has been shown to improve model generalization and performance in highly imbalanced data [10]. Additionally, the instances near the threshold are considered as potentially noisy. The alternative to labeling them as positive would be to remove them from the dataset entirely. This potentially would be removing instances with valuable information from the training data. Thus, instead of omitting the instances that the methodology is most uncertain about, they are labeled as positive as to retain as much trainable information as possible. We denote the number of instances labeled positive as P. We

evaluate our method on various P values. For the imbalanced dataset we limit P to roughly five times the expected class imbalance ratio, determined by domain expertise. We limit P so that the class label of interest does not outnumber the other class. Increasing P so that the positives outnumber the negatives would begin to resemble the way one-class classifiers are trained, in that they are best trained only on the majority class [26]. Additionally, one-class classifiers have been shown to underperform binary classifiers on highly imbalanced data [27] and for these reasons, they are out of scope of this paper. Our results show that increasing P, or labeling more uncertain instances as positive, improves supervised learning and its classification performance on original ground truth labels.

A. Experimental Datasets

We evaluate the method in this paper on two distinct cognition datasets, named Dementia Designation and Delayed Recall (DLRC). These two datasets were originally provided by the HRS [28] which has publicly available datasets for investigating work, aging, and retirement. A biennial study that began in 1992, the HRS has had more than 43,000 participants. The two datasets used were derived from an HRS dataset that is used to measure human cognition and contains data from subjects between 1992 and 2018. However, Dementia Designation and DLRC are limited to data between 2014 and 2016. The independent features of these two consist of mail-in survey data provided by the subjects. The dementia scores and delayed recall scores, which are used to generate the class labels, were collected in person by trained medical professionals. Though both datasets contain the same independent features, which are responses from the participants, and both have binary class labels, they are two distinct datasets. The participants' responses, and thus the features in the dataset, are either numeric or categorical. An example of a categorical features is the participant's answer to the question "How often are your activities during waking hours done with other people?" The available answers are "Rarely", "Sometimes", "Often", "Almost all the time", or "Uncertain, can't say". All categorical features were one-hot encoded before being used in the labeling method and before the supervised learning steps in this paper. Examples of the types of numeric responses include how many hours spent doing specified activities every week or month, respondents age, and years of education.

Autoencoders generally benefit from features that have been scaled due to the nature of the architecture and their training process. This is evident in the gradient descent optimization. Scaled features will contribute equally to the error gradient, which improves the model's convergence speed and stability during training. Unscaled features, which may have extremely different minimum and maximum values, can cause training issues. The learning algorithm can prioritize the reduction of errors of features with larger scales, as opposed to the ones with smaller scales, resulting in poorer performance. As such, we scaled the numeric features in our datasets when using the method to generate new class labels in order to best train the autoencoder. However, in order to avoid any possible data leakage in subsequent steps, we revert the scaling once all the new labels are generated and before the supervised learning steps.

TABLE I
DATASET CLASS CHARACTERISTICS

Dataset	Minority Count	Majority Count	Total Count	Minority Imbalance
Dementia Designation	148	4,311	4,459	3.319%
DLRC	2,240	2,297	4,537	49.37%

Dementia Designation and DLRC are both created for binary classification. Dementia Designation is labeled such that instances represent the respondent as either having signs of dementia or as being either cognitively normal or having some cognitive impairment that is not dementia. In this dataset, these are the minority and majority instances, respectively. DLRC, a balanced dataset, contains instances labeled as either HIGH or LOW. Instances labeled HIGH are when respondents have delayed recall scores greater than the median score of cognitively intact participants. Instances labeled LOW are when respondents have a score lower than the median score. The overall class characteristics for each of the two datasets are shown in Table I.

B. Measuring Performance

To measure the effectiveness of the synthesized labels on the balanced and imbalanced datasets, we first generate the new labels for each dataset for various levels of P. As P changes, so does the number of total instances labeled as positive. However, for each P, the total number of instances, for each dataset examined, does not change, just the positive and negative class counts. Thus, during the labeling process, no instance is omitted. We then train six supervised machine learning models on the newly created labels and then apply the models to the test set, which contains the original class labels. We train and test the models using 5-fold cross validation and repeat each experiment for 10 rounds. Each of the training folds uses the new labels and each of the test folds has the original class label for a given instance. The results shown in the tables are the average of the 50 experiments for each model. This is repeated for every level of P. It is important to note that we use the generated class labels with supervised learners solely to evaluate the performance of the new labels and the label generation method. We exclude supervised learning with original labels from our study, as our focus is on the unlabeled scenario, making the use of the original labeled data for supervised learning beyond the scope of our work.

The classification performance of the supervised models, trained on the synthesized labels, provides a good measure of overall performance of the synthesized labels, since it is a good analog to how this would be used in practice. In practice, the dataset would be entirely unlabeled, thus, our method would be used to synthesize new labels in an unsupervised way for use in supervised learning and prediction. The performance

metric of the learner on our experimental dataset would be the expected classification performance on new unseen balanced or imbalanced cognition data. In real-world scenarios, datasets often come without labels, necessitating a comparison against a baseline method that also does not rely on labels, such as an unsupervised approach. To demonstrate the viability and advantage of our method, it must outperform this baseline, thereby confirming its effectiveness.

We use the area under the precision-recall curve (AUPRC) as our performance metric. AUPRC values are derived from the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values of a classic confusion matrix. AUPRC summarizes the performance trade-off between a model's precision, or the fraction of correctly identified positive instances out of all instances predicted as positive, and recall, or the fraction of instances predicted positive out of all positive labeled instances. They are defined as: $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. We chose AUPRC as our performance metric of interest since it is a good metric for balanced data, widely used, and is a superior metric for measuring classification performance in the presence of class imbalance [29]–[31].

C. Supervised Classifiers Trained on Synthesized Labels

We measure the AUPRC performance of six supervised classifiers: Decision Tree (DT), Random Forest (RF), Extra Trees (ET), Logistic Regression (LR), Naïve Bayes (NB), and a Multilayer Perceptron (MLP). We chose these classifiers as they span a broad spectrum of machine learning paradigms which ensures a comprehensive evaluation of the synthesized labels across different model types. We aim to measure the effectiveness of the labeling technique in scenarios where data is not labeled, not to improve the supervised learners themselves. As such, we evaluate various classifiers as opposed performing hyperparameter tuning and testing various configurations of the same supervised classifier.

DTs are a commonly used supervised model that can be used for both regression and classification tasks. They are a hierarchical tree structure with a root node, branches, internal and leaf nodes. The internal nodes are conditions based on the input data and the leaf nodes are the models final numeric or categorical output. The main goal of DT is to produce a classifier that uses simple decision rules to predict a target value by using a divide and conquer training strategy. The size of the DT, determined by a model parameter, can have a large impact on performance. Since too large of a DT can lead to overfitting, pruning, or removal of nodes from the tree, is often used to reduce tree size, complexity, and improve performance. A large advantage of DT is they are relatively simple which allows for better ease of interpretability and visualization.

The second classifier we use is RF, a widely used classifier that is an ensemble of DTs. The trees are built by sampling from the training data with replacement. The ensemble aims to promote diversity among the trees and address some of DT's drawbacks. RF is easily interpretable, and each tree of the forest only considers a subset of the dataset. ET is the third classifier we use to evaluate the class labels. ET is an extension of the RF algorithm. ET chooses its split points randomly as opposed to calculating optimal split points and each tree of ET is trained on the entire dataset. Both ET and RF use majority voting from the trees to make a final output.

LR is the fourth supervised classifier and represents a more traditional statistical method that is primarily used for binary classification. It uses the logistic function curve to fit to the training data and estimates the probability an instance belongs to a class. LR uses a threshold to turn the probability into a binary output, typical 0.5. NB, the fifth classifier, is also a probabilistic classifier but is based on the Bayes' theorem. It makes the naïve assumption that a conditional independence between an instance's features and its class label exists. Lastly, MLP is a type of artificial neural network. They are made up of at least three layers, an input layer, at least one hidden layer, and an output layer, making them the most basic architecture of a neural network. MLPs, like other neural networks, can approximate non-linearly separable data, making them suitable for complex machine learning tasks. During the training phase, which uses backpropagation and a gradient descent algorithm, all training samples are used to fit the MLP to the data over several epochs.

D. Baseline Comparison

Given an unlabeled dataset, we train six supervised classifiers on our method's newly synthesized class labels to measure both the effectiveness and quality of the labels by measuring the expected performance of a learner trained on the new labels. However, it is possible to make class predictions without using class labels, as would be required for unlabeled data, by using an unsupervised method designed for anomaly detection. Thus, a performance comparison between the classification performance of our labels and that of an unsupervised anomaly detection method is needed. We use Isolation Forest (IF), originally introduced by Liu et al. in [32], as a baseline comparison. IF is a widely used outlier or anomaly detection method that is unsupervised by nature. IF is used to identify anomalous instances which are then classified as belonging to the minority, or positive group. By comparing the classification performance of IF with the classification performance of the six supervised classifiers, we can effectively measure the synthesized labels' quality and their ability to produce effective classifiers.

E. Implementation Details

We use the implementations provided by Scikit-learn [33], version 1.3.0, to define and train all six supervised classifiers and IF. We use the default values for DT and for RF we set the number of trees in the forest to 100. Our RF consists of trees with depths no larger than 4, but otherwise are the same as the DT specifications. We set the number of trees in the ET forest to 100 and the minimum number of samples per leaf to 1 and a maximum depth of 8. LR parameters are left as library defaults. We use the GaussianNB variant of NB with library defaults. For the MLP, we use one hidden

layer containing 100 neurons that uses the ReLu activation function, and the Adam algorithm for weight optimization, a type of stochastic gradient descent [34]. We use the log-loss function as the MLP's loss function. We train the MLP for 300 epochs using a constant learning rate and early stopping. IF only has one parameter, namely the contamination rate, which represents the expected class imbalance of the data, and was kept as the library default. However, it should be noted that during our study we observed that this parameter did not have any effect on threshold-based performance metrics, such as AUPRC. This is another strong indicator that AUPRC is an ideal performance metric for our work, especially when compared to non-threshold-based performance metrics, which this parameter might affect.

IV. Experimental Results

We apply the labeling technique to both the Dementia Designation and DRLC datasets. The method's input is set so that 150 positive instances are created for the Dementia Designation dataset and 1000 for DLRC. We then vary the number of positive-labeled instances for each dataset, P, such that we have a wide range of results so we can effectively measure how well the method performs across the two datasets with differing class imbalance ratios. The starting points for each are selected by domain expertise. For Dementia Designation we set P to: 150, 210, 270, 330, 390, 450, 510, 570, 63, 690, 750. For DLRC we set P to: 1000, 1150, 1300, 1450, 1600, 1750, 1900, 2050, 2200, 2350. We chose to include up to P = 750 for Dementia Designation since we observed that AUPRC stopped improving and too large of a P would be too dissimilar to the original dataset, guided by domain expertise. We chose to include no more than P = 2350 since at this point the class distribution is roughly balanced.

TABLE II
Dementia Designation - AUPRC

P	DT	RF	ET	LR	NB	MLP
150	0.1591	0.1328	0.1175	0.0877	0.1501	0.0899
210	0.1993	0.1751	0.1707	0.1361	0.2147	0.1320
270	0.2121	**0.1848**	0.1779	0.1321	0.2107	**0.1355**
330	0.2269	0.1772	0.1823	**0.1394**	0.2192	0.1321
390	0.2558	0.1728	0.1791	0.1357	0.2090	0.1302
450	0.2436	0.1614	0.1799	0.1041	0.1958	0.1037
510	0.2726	0.1707	0.1866	0.1285	0.2111	0.1139
570	0.2902	0.1710	0.1862	0.1270	0.2238	0.1172
630	0.3134	0.1713	0.1823	0.1382	0.2481	0.1274
690	0.3177	0.1713	**0.1892**	0.1384	0.2472	0.1259
750	**0.3224**	0.1696	0.1799	0.1372	**0.2507**	0.1206

It is important to emphasize that the supervised learners train on the new class labels and their performance is tested using the original class labels. This is accomplished in the 5-fold cross validation where the training folds use the synthesized labels, and the test folds use the original labels. We repeat this for 10 replications for each dataset. Table II shows the AUPRC results for all values of P for Dementia Designation and Table III shows the AUPRC results for all values of P for DLRC. Table IV shows the AUPRC scores for IF on each dataset. We train IF in an unsupervised fashion and calculate the AUPRC using the test folds using the original labels, like the supervised learners. IF also was evaluated with 5-folds of cross validation and repeated 10 times.

For each of the learners, DT, RF, ET, NB, LR, and MLP the AUPRC score is higher for the balanced DLRC dataset than the Dementia Designation, the imbalanced dataset. This can be seen for all levels of P. It is important to note that a direct comparison across P is not entirely valid since they are two distinctly different datasets. However, we can compare the two in a more general sense. DT achieves the highest AUPRC score across all learners and cognition datasets. Specifically, DT is the highest performer in Dementia Designation and in DRLC and the DT AUPRC scores are all significantly higher in the balanced DLRC dataset, as it is for the Dementia Designation. This comparison is true for all other learners, the AUPRCs are higher in the balanced data.

TABLE III
DLRC - AUPRC

P	DT	RF	ET	LR	NB	MLP
1000	0.5492	0.4676	0.4468	0.4890	0.4654	0.4962
1150	0.5564	0.4734	0.4533	0.4915	0.4853	0.5014
1300	0.5603	0.4573	0.4475	0.4787	0.4834	0.4939
1450	0.5745	0.4752	0.4494	0.4849	0.4833	0.4989
1600	0.5776	0.4532	0.4376	0.4705	0.4837	0.4882
1750	0.5906	0.4892	0.4635	0.4963	0.4992	**0.5059**
1900	0.5985	0.4809	0.4569	0.4914	0.4935	0.4999
2050	0.6126	0.4899	**0.4692**	**0.4988**	0.5060	0.5056
2200	0.6121	0.4811	0.4565	0.4919	0.5042	0.5036
2350	**0.6271**	**0.4979**	0.4676	**0.4988**	**0.5195**	0.5056

Compared to Isolation Forest on the Dementia dataset, all results outperform IF in terms of AUPRC, significantly. In some cases, such as DT, the AUPRC score is roughly 5 times greater for some P values. When looking at the DLRC results, all the learners outperform IF for most P values. ET shows the lowest scores for some P values and is similar in performance to IF until P is near its highest. As P increases, and becomes more representative of the baseline dataset, more learners outperform IF. As is the case for all datasets, DT achieves the highest AUPRC score for a given dataset and P value, and always outperforms IF. In general, AUPRC performance is higher for the balanced dataset, though for almost all learners and P values, the methodology outperforms IF in all cases.

TABLE IV
Isolation Forest Performance

Dataset	AUPRC
Dementia Designation	0.06377
DLRC	0.45715

V. Conclusion

This paper evaluates a method for synthesizing class labels in a novel context for two cognition datasets designed for identifying dementia. One dataset is balanced, and the other is

highly imbalanced. Varying levels of positive-labeled instances are evaluated across six supervised classifiers. Our selection of models provides a comprehensive evaluation of the label quality across a range of model complexities from relatively simple linear assumptions to complex nonlinear relations, and from interpretable models to ones that are harder to interpret. These classifiers produce higher AUPRC scores than the baseline unsupervised IF technique. Furthermore, we show that the labeling method produces AUPRC scores that are generally higher when the dataset is balanced than imbalanced, while still outperforming the baseline in both cases. This work not only validates the efficacy of our label synthesis method, but also shows its versatility and robustness in synthesizing high quality and effective class labels. Thus, enhancing machine learning performance in scenarios where data is initially unlabeled in the scope of both highly imbalanced and balanced data. Future work includes evaluating our method on other in other domains and exploring other levels of imbalance.

REFERENCES

[1] W. H. Organization *et al.*, *First WHO ministerial conference on global action against dementia: meeting report, WHO Headquarters, Geneva, Switzerland, 16-17 March 2015*. World Health Organization, 2015.

[2] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, "World alzheimer report 2015. the global impact of dementia: An analysis of prevalence, incidence, cost and trends." Ph.D. dissertation, Alzheimer's Disease International, 2015.

[3] W. H. Organization, "Who: Dementia." https://www.who.int/news-room/fact-sheets/detail/dementia.

[4] E. Nichols, J. D. Steinmetz, S. E. Vollset, K. Fukutaki, J. Chalek, F. Abd-Allah, A. Abdoli, A. Abualhasan, E. Abu-Gharbieh, T. T. Akram *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019," *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022.

[5] K. R. Gray, R. Wolz, R. A. Heckemann, P. Aljabar, A. Hammers, D. Rueckert, A. D. N. Initiative *et al.*, "Multi-region analysis of longitudinal fdg-pet for the classification of alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 221–229, 2012.

[6] D. Aschwanden, S. Aichele, P. Ghisletta, A. Terracciano, M. Kliegel, A. R. Sutin, J. Brown, and M. Allemand, "Predicting cognitive impairment and dementia: A machine learning approach," *Journal of Alzheimer's Disease*, vol. 75, no. 3, pp. 717–728, 2020.

[7] M. B. Ofstedal, G. G. Fisher, A. R. Herzog *et al.*, "Documentation of cognitive functioning measures in the health and retirement study," *Ann Arbor, MI: University of Michigan*, vol. 10, pp. 1 114 577 250–1 662 476 251, 2005.

[8] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE intelligent systems*, vol. 24, no. 2, pp. 8–12, 2009.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[10] R. K. Kennedy, F. Villanustre, T. M. Khoshgoftaar, and Z. Salekshahrezaee, "Synthesizing class labels for highly imbalanced credit card fraud detection data," *Journal of Big Data*, vol. 11, no. 1, pp. 1–22, 2024.

[11] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Enhancing credit card fraud detection through a novel ensemble feature selection technique," in *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2023, pp. 121–126.

[12] J. L. Leevy, T. M. Khoshgoftaar, and J. Hancock, "Iot attack prediction using big bot-iot data," *International Journal of Internet of Things and Cyber-Assurance*, vol. 2, no. 1, pp. 45–61, 2022.

[13] N. V. Chawla, "Data mining for imbalanced datasets: An overview," *Data mining and knowledge discovery handbook*, pp. 875–886, 2010.

[14] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, J. Ye, A. D. N. Initiative *et al.*, "Analysis of sampling techniques for imbalanced data: An n= 648 adni study," *NeuroImage*, vol. 87, pp. 220–241, 2014.

[15] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International journal of pattern recognition and artificial intelligence*, vol. 23, no. 04, pp. 687–719, 2009.

[16] S. Klöppel, C. M. Stonnington, J. Barnes, F. Chen, C. Chu, C. D. Good, I. Mader, L. A. Mitchell, A. C. Patel, C. C. Roberts *et al.*, "Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method," *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[17] M. Asaria, "Health care costs in the english nhs: reference tables for average annual nhs spend by age, sex and deprivation group," *CHE Research Paper*, pp. 1–25, 2017.

[18] J. R. Hodges and K. Patterson, "Semantic dementia: a unique clinicopathological syndrome," *The Lancet Neurology*, vol. 6, no. 11, pp. 1004–1014, 2007.

[19] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[20] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.

[21] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger *et al.*, *Analysis of longitudinal data*. Oxford university press, 2002.

[22] B. Seligman, S. Tuljapurkar, and D. Rehkopf, "Machine learning approaches to the social determinants of health in the health and retirement study," *SSM-population health*, vol. 4, pp. 95–99, 2018.

[23] L. Cleret de Langavant, E. Bayen, and K. Yaffe, "Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study," *Journal of medical Internet research*, vol. 20, no. 7, p. e10493, 2018.

[24] L. Cleret de Langavant, E. Bayen, A.-C. Bachoud-Lévi, and K. Yaffe, "Approximating dementia prevalence in population-based surveys of aging worldwide: An unsupervised machine learning approach," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 6, no. 1, p. e12074, 2020.

[25] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[26] J. L. Leevy, J. Hancock, T. M. Khoshgoftaar, and A. A. Zadeh, "One-class classifier performance: Comparing majority versus minority class training," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2023, pp. 86–91.

[27] J. L. Leevy, J. Hancock, and T. M. Khoshgoftaar, "Comparative analysis of binary and one-class classification techniques for credit card fraud data," *Journal of Big Data*, vol. 10, no. 1, p. 118, 2023.

[28] G. G. Fisher and L. H. Ryan, "Overview of the health and retirement study and introduction to the special issue," *Work, aging and retirement*, vol. 4, no. 1, pp. 1–9, 2018.

[29] J. L. Leevy, T. M. Khoshgoftaar, and J. Hancock, "Evaluating performance metrics for credit card fraud classification," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2022, pp. 1336–1341.

[30] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced big data," *Journal of Big Data*, vol. 10, no. 1, p. 42, 2023.

[31] J. T. Hancock III, T. M. Khoshgoftaar, and J. M. Johnson, "Using area under the precision recall curve to assess the effect of random undersampling in the classification of imbalanced medicare big data," *International Journal of Reliability, Quality and Safety Engineering*, p. 2350039, 2023.

[32] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.